UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

Studies in Anthropological Metabolomic and Microbial Approaches

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JACOB J. HAFFNER
Norman, Oklahoma
2023

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/337563

Studies in Anthropological Metabolomic and Microbial Approaches

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF ANTHROPOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Cecil M. Lewis, Jr., Chair

Dr. Jessica Cerezo-Román

Dr. Thomas Fenn

Dr. Luca Fornelli

Dr. Krithivasan Sankaranarayanan

© Copyright by JACOB J. HAFFNER 2023
All Rights Reserved.

Acknowledgements

First, I thank everyone who offered academic support and advice throughout graduate school. None of this would be possible without the support and patience of several key people. One such person is my committee chair, Dr. Cecil Lewis, whose endless optimism, patience, and wisdom enabled me to pursue this academic journey. I am sincerely grateful and appreciate mentorship – thank you. I would also like to thank Dr. Laura-Isobel McCall, who provided training in everything related to mass spectrometry-based metabolomics, including but not limited to laboratory, bioinformatics, and experiment design. Her guidance and mentorship greatly facilitated my graduate progress. I also thank Dr. Tanvi Honap and Dr. Krithi Sankaranarayanan for their contributions to my work and their assistance in brainstorming, suggestions, lessons, and their patience over my time at LMAMR, especially for training me in wet-lab protocols and bioinformatics. Additionally, I am deeply thankful to Dr. Jessica Cerezo-Román, Dr. Thomas Fenn, and Dr. Luca Fornelli for the variety of ideas, suggestions, and advice whenever we spoke, and especially for their contributions in this dissertation and broader graduate school experience.

I would also like to thank everyone at LMAMR (Nisha, Allie, Dave, Rita, Justin, Sterling, Christine, Abby, Sam, Robin, Kristen, Karissa, Nihan, Sarah, Hande, Horvey, Lizi, Cara, Amy, Chris, and everyone else) and the McCall Lab (Danya, Ekram, Mahbobeh, Azadeh, Morgan, Rebecca, Zongyuan, Jarrod, Monica, Nathan, Caitlyn, Gautham, Karina, Adwaita, Shelley, Mitchelle, and everyone else) for their patience, support, and putting up with all my ramblings/venting throughout my time at OU.

I would also like to thank all the collaborators and co-authors whom I've worked with throughout graduate school, as well as those working on any future manuscripts. Additionally, I would also like to thank everyone who contributed samples to my various research projects; none of this would be possible without your generosity and openness.

Lastly, I would like to thank my family and friends for their support and much-needed optimism (especially to offset my cynicism). To my parents, Denise and Jim, thank you for your encouragement and support throughout my time in school (and beyond). To my sister, Julie, thank you for helping keep me sane, entertained, and for all your support. Also, thank you to Lori, Caitlin, Alex, Emily, Armen, Alex "Smurf", and everyone else for your friendship and kindness.

Table of Contents

Acknowledgements iv

List of Tables and Figures x

Chapter 1 x

Chapter 2 x

Chapter 3 x

Chapter 4 xi

Chapter 5 xi

Abstract xii

Chapter 1 – Introduction 1

1.1 MOLECULAR ANTHROPOLOGY BACKGROUND 1

1.2 MASS SPECTROMETRY-BASED METABOLOMICS 2

1.3 MOLECULAR ANTHROPOLOGY AND METABOLOMICS 6

1.4 DISSERTATION STRUCTURE 8

Chapter 2 – Untargeted Fecal Metabolomic Analyses across an Industrialization Gradient Reveal Shared Metabolites and Impact of Industrialization on Fecal Microbiome-Metabolome Interactions 10

2.1 ABSTRACT 10

2.2 INTRODUCTION 11

2.3 MATERIALS AND METHODS 13

Project Design 13

Populations 14

Sample Collection 17

Ethics Approval and Informed Consent 18

LC-MS/MS Fecal Sample Preparation 19

LC-MS/MS Analysis 19

Data Analysis and Processing 21

Mmvec and Statistical Analyses 23

Data Availability 24

2.4 RESULTS AND DISCUSSION 25

2.5 CONCLUSIONS 46

Chapter 3 - Untargeted Mass Spectrometry-Based Metabolomics Reveals Responses to High Fiber Whole-Food Diets and Microbiome Composition in Mice 47

3.1 ABSTRACT 47

3.2 INTRODUCTION 49

3.3 MATERIALS AND METHODS 51

Experiment Design 51

Dietary Formulation 53

Gnotobiotic husbandry 53

Colonization of GF mice with human fecal samples and diet interventions 53

Metabolite extraction for untargeted metabolomics 54
Liquid Chromatography-Tandem Mass Spectrometry Data Acquisition 55
Untargeted Metabolomics Data Analysis 55
Data availability 58
3.4 RESULTS AND DISCUSSION 59
Diet and gut community differentially modulate cecal metabolites 59
Diets and gut communities modulate serum metabolites 70
3.5 CONCLUSIONS 71
Chapter 4 - Mass Spectrometry Database of Archaeologically Relevant Plants 73
for Organic Residue Analysis 73
4.1 ABSTRACT 73
4.2 INTRODUCTION 74
4.3 MATERIALS AND METHODS 78
Project Design 78
Sample Collection 79
Sample Treatment and Extraction 80
LC-MS/MS Analysis 83
Data Analysis and Processing 84
Data Availability 85
4.4 RESULTS AND DISCUSSION 85

4.5 CONCLUSIONS 91

CHAPTER 5 - CONCLUSIONS 93

Bibliography 96

Supplementary Material A 114

Author List and Affiliations 114

Author Contributions 114

Acknowledgements 115

Supplementary Figures 2.1-2.4 116

Supplementary Tables 2.1-2.4 174

Supplementary Material B 240

Author List and Affiliations 240

Author Contributions 240

Acknowledgements 241

Supplementary Figures 3.1-3.4 242

Supplementary Material C 248

Author List and Affiliations 248

Author Contributions 249

Acknowledgements 249

Supplementary Table 4.1 250

List of Tables and Figures

Chapter 1

No tables or figures

Chapter 2

Table 2.1: Sampled population metadata. 14

Figure 2.1: Fecal metabolomic profiles follow an industrialization gradient. 28

Table 2.2: Top 30 most differential metabolite features as determined by random forest classifier. 36

Figure 2.2: The shared human metabolome. 39

Figure 2.3: Principal component analyses (PCA) illustrate probable metabolite-microbe co-occurrences. 43

Chapter 3

Figure 3.1: Experiment design. 51

Table 3.1: Composition of Teklad global 18% protein rodent diet. 52

Figure 3.2: Gut microbial community and diet modulate cecal metabolites. 60

Figure 3.3: Unique and overlapping metabolite features with differential abundance between donor and diet combinations in cecal contents (UpSet plot). 64

Figure 3.4: Plant-derived metabolite families differentially impacted by microbiome and diet. . 65

Figure 3.5: Impact of donor and diet on bile acid relative abundance. 67

Figure 3.6: Impact of donor and diet on commendamide and kynurenic acid. 69

Chapter 4

Figure 4.1: Plant sampling locations. 78

Figure 4.2: Experiment design. 82

Figure 4.3: Molecular networks highlight region-based abundances of plant chemical compounds. 88

Chapter 5

No tables or figures

Abstract

Human biology is shaped by layers of interacting biomolecules, from genes to proteins and small molecules called metabolites. These molecules are crucial participants in biochemical reactions, representing biological information within and above the genome. Investigating the system of small molecules (the metabolome) offers deep inquiries into the interplay between human health, biology, behavior, and the environment. Anthropological interest in these small molecules is growing, but more work is needed to integrate metabolomics within anthropology. In this spirit, this dissertation applies metabolomics approaches with mass spectrometry to address anthropological questions. Chapter one provides a background over mass spectrometry, metabolomics, and anthropological engagement with these approaches. From there, chapter two directly applies mass spectrometry-based metabolomics to study human fecal metabolomes of individuals practicing varying lifeways from Africa and the Americas, revealing key chemical components to the human gut. Next, chapter three investigates mouse metabolites modulated by gut microbes responding to three high-fiber high-fat diets, revealing personalized health and biological responses to dietary shifts. Chapter four then applies mass spectrometry-based metabolomics to preserved residues found in archaeological contexts by analyzing plant and dietary materials relevant to archaeological communities. We discuss new developed methods to analyze these samples, report plant molecules associated with these samples, and detail plans to make these data publicly available for archaeologists and other scholars to integrate into their analyses, improving results generated from these archaeological samples. Together, these three projects explore novel methods of integrating these metabolomic approaches and data into anthropological investigations of past and present human health, behavior, and biology.

Dissertation Keywords: Metabolomics; Microbiome; Mass spectrometry; Residue analysis; Molecular anthropology

Chapter 1 – Introduction

1.1 MOLECULAR ANTHROPOLOGY BACKGROUND

Knowledge of the molecular dimensions to human health and biology have grown rapidly over recent years due to analytical and technological advancements enabling deep inquiries into systems biology. These approaches revealed how genetic information flows from genes into transcribed RNA, which are then translated into proteins using small molecules to participate in chemical reactions along molecular pathways[1]. In anthropology's investigation of the human experience[2], subdisciplines like biological anthropology adopted these analytical approaches. Coupled with the growth of fields like genomics and systems biology, biological anthropology expanded molecular anthropology to represent a series of applied chemical, biological, and microbiological approaches to answering anthropology questions[3].

Recent studies in molecular anthropology have had a wide range of impact, from understanding human-ape evolutionary relationships through comparing microbial profiles between humans and great apes[4], illustrating new ways of exploring health states and disease susceptibilities[5,6], studying Neanderthal DNA[7,8], analyzing ancient proteins reflecting past human diets[9], to name a few. Using cutting-edge genomics, molecular anthropology research has extensively characterized the genes, transcripts, and proteins involved in human biology, the classic "central dogma" of biology applied to humans. However, above these "omic" systems is metabolomics. Far less anthropological research has comparatively focused on the small molecules responsible for biological processes. These small molecules, called metabolites, represent biological information beyond the genome as they shape the biological phenotype[1,10,11].

Thus, anthropological understandings of human biology and health, especially as they relate to unique lifestyles and environments, require knowledge of these metabolite-oriented components.

1.2 MASS SPECTROMETRY-BASED METABOLOMICS

While many definitions exist within the field, small molecules (metabolites) refer to any molecule involved in some biochemical reaction <1500 Daltons (Da) in size[1,12]. Together, the sum of metabolites present in a biological system is called the metabolome[1,13]. Importantly, the metabolome not only includes endogenous molecules produced by the host (and their associated microbes), but also exogenous molecules accrued from the environment[12,14]. These molecules represent the intermediate and end products of any given biochemical reaction[1,11]. Metabolomics provides a snapshot of the biological state of an organism/sample, which combined with gene, transcript, and protein abundances, constitute the molecular phenotype of an organism. For humans, the metabolome represents the direct interaction between host genetics and the environment, cultural practices, lifestyles, etc. down to a molecular level.

One common approach to characterize metabolomes is through mass spectrometry (MS). Due to the extensive variability of metabolites, it is impossible to completely characterize the metabolome[12]. Thus, researchers choose MS instruments, parts, and any laboratory protocols based on their specific research questions. While these details vary between projects, MS detects, quantifies, and analyzes the molecules present in biological samples by sorting the detected molecules according to their masses[15,16]. Four critical steps are involved in MS analysis: 1. sample molecules are introduced to the MS instrument; 2. sample molecules become ionized to create charged particles; 3. charged ions are separated based on their properties as they pass through the instrument (with optional steps to filter out certain molecules); 4. molecules of interest are quantified by a detector[11,15]. Some MS techniques include an additional step called

tandem mass spectrometry (MS/MS, or MS2), which fragments charged ions after detection and runs these charged fragments through the detector again[15]. This second fragmentation step improves molecule annotation by providing more detailed structural information about the sample molecules. All MS instruments and components complete the four major steps of introduction, ionization, separation, and detection in different ways and each variation introduces different biases[15]. In the end, these steps provide detailed data about the molecular composition of biological organisms/samples.

MS-based metabolomic research generally falls into one of two categories: targeted or untargeted. Targeted analyses focus on quantifying known or suspected molecules in a sample or biological system[14,17], while untargeted analyses characterize the global range of detectable metabolites in a sample or system[15,16]. Since no approach completely captures all aspects of a metabolome, researchers will select the targeted or untargeted approach, metabolites of interest, experimental conditions, MS instrument, and analysis techniques based on what best addresses their specific research question. Moreover, it is important to note analytical challenges are especially prevalent when analyzing untargeted data as it focuses on molecules that often have minimal information beyond mass[18]. In other words, while an untargeted analysis can identify an informative trait of the molecule like mass, there is much about its nature that is not fully understood. Identifying and analyzing patterns of unknown metabolites present unique challenges for untargeted metabolomics, but these analytical tools are rapidly improving over time.

Targeted analyses are more commonly practiced than untargeted, partly because targeted data results in fewer data points (metabolites of interest) than untargeted analyses[18]. Indeed, targeted approaches are especially frequent in anthropology studies compared to untargeted, partly due to

the anthropological questions being addressed are simpler and more focused on a specific molecule of interest, like molecules associated with tobacco or chocolate. However, as discussed later, these targeted approaches also have limitations and create assumptions and biases. An ideal future for most anthropology questions is one in which an untargeted approach could be used effectively. For this dissertation, untargeted metabolomics will be the primary approach due to its capacity to reveal metabolites with fewer researcher assumptions.

Another major divergence among MS techniques involves sample introduction through different types of chromatographic separation, namely gas chromatography (GC) and liquid chromatography (LC)[15,19]. While many techniques for chromatographic separation exist, GC and LC are among the most common separation protocols[19], especially for anthropology[20–24]. Chromatographic separation separates sample molecules according to their properties by attaching samples to a stationary phase (usually via a chromatography column) and use a mobile phase to pass samples through the stationary phase[15,19]. Molecules will pass through these phases differently, according to factors like matrix composition, molecular weight, absorption affinity, and more[19]. GC and LC primarily differ based on the state of mobile phase used, with GC using a gas and LC using a liquid mobile phase[15,19]. Another difference between GC and LC is that GC vaporizes sample molecules for the mobile phase to interact, whereas LC needs no vaporization[15,19]. Stationary phase selection, mobile phase used, vaporizing component (or lack thereof), etc., all influence the types of molecules introduced to the MS instrument. Ultimately, GC works best for thermally stable molecules, is biased towards nonpolar molecules, and generally works well with metabolites up to 600 Da in size, while LC is best-suited for thermally volatile, polar molecules up to 1500 Da[15,19]. Due to its wider detection thresholds, LC is preferable for untargeted metabolomics, but in anthropological literature, but GC is more

common despite its narrower resolution window compared to LC[23,25–28]. Nonetheless, researchers select between these approaches based on their molecules of interest and whether targeted or untargeted analyses are employed.

While MS-based metabolomics is relatively new compared to the other major omics fields, it has emerged as a prosperous and distinct field covering a wide range of topics. Common targeted analysis topics include drug or natural product discovery[29–31] and exploring metabolic pathways like those involved in drug responses[1,14]. Additionally, untargeted analyses commonly explore sample spatial components to molecules and metabolomes[32–34], examine metabolomic responses to stimuli like disease/health responses[35–38], or provide chemical inventories of sample materials[39,40]. Beyond the wealth of research topics, metabolomics scholars also developed field standards for generating and reporting MS-based metabolomics data, resulting in the metabolomics standards initiative (MSI)[41]. The MSI criteria and overall goals aim to maximize the utility of MS data by encouraging full reports about all details involving metabolite detection, quantification, and identification like laboratory protocols and experiments, sample collection and treatment, MS analysis, and data analysis. One particularly important MSI topic is the levels of criteria for molecular identification, also known as annotation. Briefly, the MSI authors discuss four levels of annotation, with level four representing lowest-confidence annotations and level one as highest-confidence annotations, based on how the molecule is identified[41]: level four: unknown compounds that are differentiated by spectral properties; level three: identified by chemical class or spectral similarity to known chemical classes; level two: annotations based on database library references; and level one: annotations based on matches to authentic chemical standards[41]. The MSI guidelines have become the golden standard for conducting MS-based metabolomics studies and defined subsequent MS-based metabolomics research[38,42–44].

Additionally, new MS-based metabolomics technical and analytical tool advancements like the development of platforms like Global Natural Products Social Molecular Networking system[45], molecular networking techniques[46,47], creating new metabolomic databases[44,46], etc., enabled the field to grow rapidly in quality and breadth of research. As more new tools and techniques come out, MS-based metabolomics will only improve.

1.3 MOLECULAR ANTHROPOLOGY AND METABOLOMICS

Broadly, anthropology has used metabolomics to investigate various aspects of human evolution, health, disease, and cultural behavior. In biological contexts, metabolomics studies reveal how aspects of our unique human lifestyles and behaviors, like diet or industrialization patterns, interact with our genetics and molecular biology[43,48,49]. Furthermore, these anthropology studies have also examined metabolomes of traditional foragers[50], compared oral metabolomes of ancient and modern humans[51], developed forensic methods through integrating metabolomic data with other multi-omic analyses[52], and investigated medicinal plant consumption by Neanderthals[53]. Another approach involves the microbiome, the collection of microbes and their genetic material[54], and the metabolic pathways associated with these microbes. This microbiome-metabolome connection is prevalent in many case studies, especially anthropological metabolomics studies, to explore the functional consequences of our residential microbes. Examples include identifying novel bile acids modified by gut microbiota[30], characterizing traditional human fecal metabolomes by association to host microbiome[50,55], and comparing ancient and modern human oral microbiome profiles with integrated metabolomics data[56]. These studies made strides towards better integrating MS-based metabolomics with biological and molecular anthropology, but there is still capacity for growth.

The bulk of anthropological MS-based metabolomics studies involve organic residue analysis, a common archaeological technique. Residue analysis investigates the biomolecules found in organic residues preserved on archaeological samples. The detection and chemical identity of these molecules on archaeological objects provide information about past human cultural behavior[21,57]. While a wide variety of techniques for residue analysis exist, metabolomics-like analyses using mass spectrometry are the most common[21,23,58]. Such metabolomics-like residue analyses reported the molecular components of preserved residues, such as identifying botanical origins for historic ritual drinks[59], characterizing ancient wine residues and potential source regions[22,60], describing chemical components of ancient Egyptian canopic jars[20], and detecting nicotine in smoking pipe fragments[61,62]. While archaeology has used MS-based metabolomics approaches via residue analysis for decades, the explicit association with metabolomics is relatively new[20,62–64].

By investigating the small molecules intrinsic to any biochemical processes, MS-based metabolomics is well-poised to address anthropological questions. Characterizing human metabolomes through MS ultimately reveals human biology down to a functional, chemical level[1,65]. Moreover, MS-based metabolomics is uniquely suited to examine how the environment interacts with and subsequently informs human biology, allowing anthropologists to study the biological consequences of our unique lifestyles and behaviors. In health contexts, metabolomic investigations of diseases with multidimensional causes, like diabetes, can illustrate the suite of molecules acquired from different sources like genetic predisposition, diet, lifestyle, and more [66,67]. This biology-environment interplay as gleaned through MS-based metabolomics even extends to studying past human cultural behaviors, such as through residue analysis, and how these past behaviors influenced past human health. Thus, MS-based metabolomics brings much

to any anthropological explorations of human biology and its relationships with health, environments, and behavior, in both past and/or present contexts.

1.4 DISSERTATION STRUCTURE

This dissertation highlights how MS-based metabolomics can address anthropological questions by describing three studies using MS-based metabolomics in anthropological contexts. These case studies are represented by chapters 2-4, while chapter 5 is a conclusions section.

Chapter two uses MS-based metabolomics to examine the fecal metabolomes of humans from Africa and the Americas representing ranges of human lifestyle diversity, including western industrial communities and traditional hunter-gatherers, with a focus on industrialization. This novel emphasis on lifestyle and industrialization varieties offers a unique exploration into how lifestyle and behavior influence human biology down to a chemical level. Chapter 2 also reveals the core chemistry of the human gut and potential connections between this shared gut chemistry and the gut microbiome.

Chapter 3 incorporates MS-based metabolomics approaches to investigate metabolomic alterations by gut microbial responses to different high fat and high fiber combination diets. This project focuses on mice metabolomes to represent a controlled setting for exploring microbiome-metabolome consequences of dietary shifts. In this chapter, we reveal personalized health responses to these dietary shifts by examining metabolites directly influenced by gut flora. These results represent new approaches for examining dietary effects on health and biology.

Chapter 4 studies plant and dietary samples relevant to archaeological communities from different regions in the Americas. These data represent raw dietary materials associated with archaeological metabolomics-like investigations of organic residues preserved on archaeological

samples. In this project, we describe an untargeted metabolite extraction and analysis method to provide broad metabolomic profiles of these samples, study abundances patterns of plant-associated molecules, and discuss plans for making these data publicly accessible.

Combining these chapters together, this dissertation shows the value of integrating MS-based metabolomics with anthropology. There is certainly room for growth in applying metabolomics to anthropology, but this dissertation represents steps towards that growth by focusing on key case studies and new methods that can help expand anthropological knowledge about molecular components of human health, biology, behaviors, and culture.

Chapter 2 – Untargeted Fecal Metabolomic Analyses across an Industrialization Gradient Reveal Shared Metabolites and Impact of Industrialization on Fecal Microbiome-Metabolome Interactions[1,2]

2.1 ABSTRACT

The metabolome is a central determinant of human phenotypes and includes the plethora of small molecules produced by host and microbiome or taken up from exogenous sources. However, studies of the metabolome have so far focused predominantly on urban, industrialized populations. Through an untargeted metabolomic analysis of 90 fecal samples from human individuals from Africa and the Americas—the birthplace and the last continental expansion of our species, respectively—we characterized a shared human fecal metabolome. The majority of detected metabolite features were ubiquitous across populations, despite any geographic, dietary, or behavioral differences. Such shared metabolite features included hyocholic acid and cholesterol. However, any characterization of the shared human fecal metabolome is insufficient without exploring the influence of industrialization. Here, we show chemical differences along an industrialization gradient, where the degree of industrialization correlates with metabolomic changes. We identified differential metabolite features such as amino acid-conjugated bile acids and urobilin as major metabolic correlates of these behavioral shifts. Additionally, coanalyses with over 5,000 publicly available human fecal samples and cooccurrence probability analyses with the gut microbiome highlight connections between the human fecal metabolome and gut microbiome. Our results indicate that industrialization significantly influences the human fecal

[1]Adapted from Haffner et al. (2022). Untargeted Fecal Metabolomic Analyses across an Industrialization Gradient Reveal Shared Metabolites and Impact of Industrialization on Fecal Microbiome-Metabolome Interactions". *Msystems* 7(6), 300710-22. (https://journals.asm.org/doi/10.1128/msystems.00710-22)
[2] Full author list available in Supplementary Material A.

metabolome, but diverse human lifestyles and behavior still maintain a shared human fecal metabolome. This study represents the first characterization of the shared human fecal metabolome through untargeted analyses of populations along an industrialization gradient.

2.2 INTRODUCTION

Metabolites fit as the final stage of biology's central dogma: DNA transcribed into RNA translated into proteins which enzymatically interact, form, and shed into small molecules as part of the biochemical pathways of metabolism[1,11,13]. For this study, we define a metabolite as any small molecule (<1,500 Da) involved in biochemical pathways and the metabolome as the collection of these small molecules within a biological system[1,12,29]. Using the definition from the Human Metabolome Database, these endogenous metabolites (synthesized by the host or their resident microbiota) are supplemented by exogenous small molecules acquired from external sources, such as cosmetics, medication, dietary sources, and pollution[44]. The human metabolome thus contains both endogenous and exogenous metabolites, representing the nexus of genetic and environmental influences[29,68–70].

Characterizing the fecal metabolome requires an understanding of how it is influenced by different factors, such as industrialization[71,72]. Broadly, industrialization is a series of economic and technological changes relating to the processing and distribution of resources that ultimately cause a shift from agrarian to industrial societies[73,74]. Such changes generally involve an increase in manufactured products compared to agriculture/hunting and other raw products, a greater percentage of workers employed in industrial workplaces over agriculture, and changes in the physical landscape such as increased construction of built environments[75,76]. Industrialization is often linked with urbanization, which refers to social and demographic shifts increasing population size and density within a settlement[75]. These processes lead to industrialized-urban

populations exhibiting denser populations[75], reduced exposures to nature-derived molecules but increased exposure to human-derived molecules[33,43,77–79], an indirect relationship with food sources[80,81], and dietary shifts[49,81] compared to non-industrial rural populations. Moreover, industrialization is associated with significant biological changes, such as reduced microbial diversity[79,82–84], increased allergic diseases[85,86] and asthma[87], and heightened susceptibility to illnesses such as inflammatory bowel disease[88–90], although further work is required to definitively show industrialization processes as the primary cause of these changes given that such health conditions have complex causes[91,92]. Investigations into industrially-caused fecal metabolomic shifts have identified differences based in amino acids, amines, sphingolipids, and hexoses, among others[48,49,92]. Some studies detailed human fecal metabolomes by comparing rural and urban populations and found differences in levels of acylcarnitines, amino acids, and short-chain fatty acids[48,50,93]. However, such studies employed targeted/semi-targeted metabolomic approaches and/or sampled a single human population[48–50,83,93]. As a result, these studies do not represent ranges of human diversity and behavior, highlighting the need for broader investigations of the human fecal metabolome in terms of geographic range and chemical space.

We performed untargeted liquid chromatography mass spectrometry (LC-MS)-based metabolomics on 90 fecal samples obtained from six human populations from diverse geographic regions (Figure 2.1a; Table 2.1; Table S2.1). These populations included male and female children and adults. Our sampled populations were categorized corresponding with their degree of industrialization, based on lifestyle factors such as dietary practices, built environment, population, etc. (see Materials and Methods for further details on categorization). Importantly, we included two populations with similar degrees of industrialization but from distinct

continents, to control for any geographic confounders. This key aspect had not been considered in prior industrialization-focused metabolomics research. Our populations include: Norman (USA; urban industrial; 18 samples); Guayabo (Peru; rural industrial; 12 samples); Tambo de Mora (Peru; rural industrial; 14 samples); Boulkiemdé (Burkina Faso; rural traditional; 11 samples); Tunapuco (Peru; rural traditional; 24 samples); and Matses (Peru; isolated traditional; 11 samples).

2.3 MATERIALS AND METHODS

Project Design

Fecal samples from six human populations were analyzed, representing ranges of industrialization. Populations were categorized to reflect varying degrees of lifestyle behavior and industrialization, based on diet, access to pharmacies and public markets/stores, housing structure, and population density[43]. Grouped categories are: urban industrialized (highly industrial urban population, typical Western industrialized population; *n*=18); rural industrialized (industrialized population, but primarily rural environment over urban; *n*=26); rural traditional (rural communities with some industrialization; *n*=35); and isolated traditional (isolated rural community with little to no industrialization influence; *n*=11). The study populations include: Norman (*n*=18), Oklahoma, USA, a standard Western industrialization population located in the Oklahoma City metropolitan area; Guayabo (*n*=12), Peru, a large rural town influenced by industrialization; Tambo de Mora district (*n*=14), Peru, a large rural district influenced by industrialization; Boulkiemdé province (*n*=11), Burkina Faso, with some industrialization influence; Tunapuco (*n*=24), a traditional rural community located in the Andean Highlands with minimal industrialization influence; and the Matses (*n*=11), an isolated traditional hunter-gatherer community from the Peruvian Amazon (Figure 2.1; Table 2.1; Table S2.1). All

populations contained both males and females of ages three to 77. Individuals under the age of three were excluded from analyses because gut microbiomes do not stabilize and resemble adult microbiomes until after age three[94,95]. Dietary information was collected through standardized questionnaires and personal interviews with the sample donors. Thus, all diet information was self-reported based on these questionnaires and interviews.

						Age distribution			Sex	
Population	**Abbreviation**	**Geographic Origin**	**Industrialization Group**	**Sample size (n)**	**Time Kept on Ice Before Frozen**	**5-17 years**	**18-44 years**	**45+ years**	**Female**	**Male**
Total				90		28	47	15	47	29
Norman	NO	Norman, Oklahoma, United States	Urban Industrial	18	Within 24 hours	0	18	0	7	11
Guayabo	GU	Guayabo, Peru, South America	Rural Industrial	12	Within 4 days	4	4	4	8	0
Tambo de Mora	TM	Tambo de Mora District, Peru, South America	Rural Industrial	14	Within 4 days	5	7	2	7	1
Boulkiemdé	BF	Boulkiemdé Province, Burkina Faso, Africa	Rural Traditional	11	Within 24 hours	0	6	5	6	5
Tunapuco	HCO	Andean Highlands, Peru, South America	Rural Traditional	24	Within 4 days	13	9	2	13	7
Matses	SM	Peruvian Amazon, South America	Isolated Traditional	11	Within 4 days	6	3	2	6	5

Table 2.1: Sampled population metadata.

Populations

Fecal samples from Norman, Oklahoma, USA, were analyzed for this project (*n*=18), representing western industrial lifestyles and diets. Norman residents live in the Oklahoma City metropolitan area, exemplifying a highly industrialized environment. Self-reported diets

generally consisted of regular dairy consumption plus processed and/or prepackaged foods like canned vegetables[82]. Additionally, regular meat consumption was common to Norman individuals compared to our other sampled populations. Due to the strongly industrialized setting and greater consumption of meat, dairy, and processed food products, this population was categorized as urban industrial.

We also selected fecal samples from the Guayabo (*n*=12) and Tambo de Mora (*n*=14) populations, which practice similar lifestyles. These populations exhibit rural lifestyles and diets but are still strongly influenced by industrialization. Both communities have regular access to public markets and pharmacies and live in densely packed areas. Their diets are generally reliant on foods obtained from these markets, as well as local produce and livestock. While the Guayabo diet commonly consists of maize with some meat and dairy consumption, the Tambo de Mora population relies more on fish, due to their proximity with the Peruvian coastline. Consumption of processed foods is common for both communities, albeit less so than Norman individuals. Because the Guayabo and Tambo de Mora communities exhibit some characteristics of non-industrial and industrial lifestyles and live in primarily rural settings, these populations were categorized as rural industrial.

The Boulkiemdé (*n*=11) and Tunapuco (*n*=24**)** communities represent the next degree of industrialization in our sampled populations. Although these populations are from Africa and South America, respectively, they practice similar traditional non-industrial, rural lifestyles and share some features of industrialized populations such as access to public markets. The Boulkiemdé samples were collected from the Boulkiemdé province of Burkina Faso. This Burkinabé community practices an agricultural lifestyle, usually growing their own crops, raising livestock, and with rare dairy consumption. Boulkiemdé meat consumption often ranged from

once every 1-3 weeks or once every 4-6 months. Vegetable consumption was high in self-reported diets by Boulkiemdé individuals. Common vegetables included cabbage, okra, eggplant, beans, carrots, potato, manioc, couscous, bread, rice, corn, etc. Processed foods like canned vegetables were highly rare. Meanwhile, the Tunapuco population have similar traditional agricultural lifestyles, relying on local produce and livestock[82]. Residing in the Peruvian Andes highlands, the Tunapuco people have diets largely consisting of root and stem tubers, bread, and rice[82]. The Tunapuco people occasionally consume animal proteins and dairy products such as cuy, beef, pork, or sheep**.** Overall, rice, mote, carrots, cabbages, bread, cuy, oca, and potatoes (fermented, dehydrated, etc.) are the most common self-reported foods for the Tunapuco people. Additionally, Tunapuco residents have access to lowland markets, which offer other dietary sources like fruit (apples, bananas, pineapples, mangos, etc.), depending on seasonal availability[82]. Similar to the Boulkiemdé community, processed foods are rarely consumed by the Tunapuco people. Since both the Boulkiemdé and Tunapuco communities sampled for this project lived in largely rural yet partly industrial environments with diets focused more on raw food products, these populations were grouped as rural traditional.

Our last sampled population is the Matses (n=11). The Matses people practice traditional hunter-gatherer lifestyles[82], making them unique for this study. Their diet is based heavily on tubers, plantains, fish, and game meat. Specifically, varieties of manioc, plantains/bananas, and fish are staples of the Matses diet, while bushmeat, reptiles, birds, bread, and other crops are less frequent. Dairy and processed foods are very rarely consumed by the Matses community. Due to their location in the Amazonian regions of Peru and unique lifestyles characterized by self-reliant food production over processed foods, the Matses are almost completely isolated from external

sociocultural and economic influences like industrialization[82], so they were categorized as isolated traditional.

Sample Collection

Fecal material was deposited into polypropylene containers and then put on ice. Samples were kept in ice while in the field until arriving at research facilities equipped with freezers. The Norman samples were kept in ice after collection and frozen at the laboratory within 24 hours. The Peruvian samples were secured similarly to the Norman samples. After collection, samples were stored on ice for four days until arriving at Lima, Peru. Samples were frozen and sent to the laboratory in Norman, Oklahoma.

The Norman, Tunapuco, and Matses samples had previously been aliquoted and underwent 16S rRNA gene sequencing for an earlier study[96], using the MoBio PowerSoil DNA Isolation Kit protocol (full details can be found in the original article[96]). The raw fecal samples were otherwise kept frozen at -80 °C until use for this project.

Boulkiemdé samples were collected similarly to Norman and Peruvian samples. After collection, Boulkiemdé samples were frozen at -20 °C within 24 hours and kept frozen overnight. Samples were thawed the following evening to extract DNA, refrozen at -20 °C, and kept frozen until shipped to the laboratory in Norman, Oklahoma. Upon arrival, 2 g of fecal material was extracted from each sample for anaerobic culturing. Following this 2 g aliquoting, samples were frozen at -80 °C until use for this project.

While field conditions mandated different storage protocols, we confirmed that these effects are overshadowed by the industrialization gradient (see 2.4 Results and Discussion).

Full metadata with health conditions, such as primary water sources, pharmaceutical consumption, date of latest hospital visit, etc., were collected for the Boulkiemdé samples. However, Norman and Peru samples had been collected several years before the Boulkiemdé samples, and unfortunately lack similar detailed metadata about health conditions. While this metadata cannot be provided for Norman and Peru samples, the full de-identified metadata for the Boulkiemdé samples are available upon request.

Ethics Approval and Informed Consent

Norman and Peruvian sample collection was conducted under the supervision of the University of Oklahoma and the Ethics Committee of the Peruvian National Institute of Health Ethical protocols for community engagement and sample collection were developed through collaboration with representatives and authorities from each sampled region and in accordance with institutional regulations. Norman samples were collected through informed consent from local volunteers recruited from posted flyers or word-of-mouth. All Peruvian samples were obtained through community engagement with local and national authorities and informed consent with consultation from the Center for Intercultural Health of the Peruvian Institute of Health and Peruvian National Institute of Health ethics committee. This project was reviewed and approved by the research ethics committee of the Instituto Nacional de Salud del Peru (Projects PP-059-11, OEE-036-16). These approvals extended to any future molecular analyses (i.e., MS-based metabolomics) and were done with approval of the stakeholders and communities.

Human fecal samples were collected with informed consent from resident volunteers in central Burkina Faso under the ethics review committee of Centre MURAZ, a national health

research institute in Burkina Faso (IRB ID No. 31/2016/CE-CM). OU IRB deemed this project consistent with US policy 45 CRF 46.101(b) exempt category 4 (OU IRB 6976).

LC-MS/MS Fecal Sample Preparation

The sample preparation protocol used for this project was adapted from a global metabolite extraction protocol with proven success[35,97]. Samples were thawed and 500 µl of chilled LC-MS grade water (Fisher Scientific) was added to 50 mg of fecal material. Next, a TissueLyzer homogenized samples at 25 Hz for three minutes. Following homogenization, chilled LC-grade methanol (Fisher Scientific) spiked with 4 µM sulfachloropyridazine as the internal standard (IS) was added, bringing the total concentration to 50% methanol. The TissueLyzer homogenized samples again at 25 Hz for three minutes, followed by overnight incubation at 4 °C. The next day, samples were centrifuged at 16,000 x g at 4 °C for ten minutes. Aqueous supernatant was then removed and dried using a SpeedVac vacuum concentrator. Dried extracts were frozen at -80 °C until the day of MS analysis. Immediately prior to MS analysis, extracts were resuspended in 150 µl chilled LC-MS methanol:water (1:1) spiked with 1 µg/ml sulfadimethoxine as a second IS. After resuspension, samples were diluted to a 1:10 ratio. Diluted samples were sonicated using a Fisher Scientific Ultrasonic Cleaning Bath at maximum power for ten minutes. Supernatants were spun briefly to remove any particulates, then loaded into a 96-well plate for MS analysis. One well contained only 150 µl of the resuspension solution to serve as blank control.

LC-MS/MS Analysis

LC was performed on a ThermoFisher Scientific Vanquish Flex Binary LC System with a Kinetex C18 core-shell column (50 x 2.1 mm, 1.7 µM particle size, 100 Å pore size). LC column was kept at 40 °C and the sample compartment was held at 10 °C. The LC System was coupled

to a ThermoFisher Scientific Q Exactive Plus Hybrid Quadrupole-Orbitrap Mass Spectrometer for MS/MS analysis. For the LC mobile phase, Solvent A was LC-MS grade water (Fisher Scientific) with 0.1% formic acid and Solvent B was LC-MS grade acetonitrile (Fisher Scientific) with 0.1% formic acid. Elution gradient started at 5% Solvent B for one minute, increased to 100% Solvent B until minute nine, held at 100% Solvent B for two minutes, dropped to 5% Solvent B over 30 seconds, and 5% Solvent B for one minute as re-equilibration. Samples were injected in random order with an injection volume of 5 µl. After elution, electrospray ionization was conducted with spray voltage of 3.8 kV, auxiliary gas flow rate of 10, auxiliary gas temperature at 350 °C, sheath gas flow rate at 35, and sweep gas flow at 0. Capillary temperature was 320 °C and S-lens RF was 50 V.

MS1 scan range was 100-1,500 *m/z*, MS1 resolution was set to 35,000 and MS1 AGC target to 1e6. MS1 data were obtained in positive mode and MS2 data were obtained using data-dependent acquisition. In each cycle, 5 MS/MS scans of the most abundant ion were recorded. Both MS1 and MS2 injection times were set at 100 ms. MS2 resolutions were set to 17,500, MS2 AGC target was set to 5e5, and the inclusion window to 2 *m/z*. MS/MS was conducted at an apex trigger of 2-8 seconds and an exclusion window of 10 seconds. MS/MS collision energy gradually increased from 20-40%.

Authentic standards also underwent LC-MS/MS analysis to validate metabolite annotations. A total of 15 standards were purchased from AA Blocks (hyocholic acid, 13-docosenamide), AvaChem (lenticin), Biosynth (bilirubin, N-acetylmuramic acid, fructosyl-L-lysine), BLD Pharm (N-palmitoylglycine, trans-ferulic acid), ChemScene (leucine enkephalin), LGC Standards (L-saccharopine), Sigma-Aldrich (L-abrine, N-acetyl L-phenylalanine, enoxolone, octadecanamide, lithocholic acid, paraxanthine), and VWR (nicotinamide N-oxide).

Each pure standard was diluted to 100 μM, 50 μM, 10 μM, 5 μM, and 1 μM concentrations. All standards (and their five dilutions) were analyzed according to the same LC-MS/MS parameters as the original samples. Additionally, fecal extracts with the highest abundance for each standard were re-analyzed as part of the same LC-MS/MS batch to ensure standard peaks were present in samples and to prevent confounding from retention time shifts caused by the gap between initial data acquisition and annotation validation.

Data Analysis and Processing

MSConvert v3.0.19014[98] converted raw data files to mzXML format in preparation for data processing via feature-based molecular networking (FBMN)[99]. MZmine v2.33[100] was used to identify MS features for all samples (Table S2.4). All non-gap filled analyses were performed using identical parameters to the gap filling steps in MZmine, with the exception of the gap filling step. After feature filtering, only features with abundance three times greater than the abundance of blanks were retained in these analyses. Total ion current (TIC) normalization was conducted through R programming language v3.5.3[101] in Jupyter Notebook[102]. FBMN and library spectral database searches were completed using the FBMN workflow on Global Natural Products Social Molecular Networking (GNPS)[45]. FBMN GNPS parameters for MS/MS analysis were as follows: precursor and fragment ion mass tolerance: 0.02 Da, minimum cosine score for networking and library matches: 0.7, minimum number of matched MS2 fragment ions for networking and library matches: 4, network topK: 50, maximum connected component size: 100, maximum shift between precursors: 500 Da, analog search: enabled, maximum analog mass difference: 100 Da, precursor window filtering: enabled, 50 Da peak window filtering: enabled, normalization per file: row sum normalization. Results were analyzed by visually evaluating mirror plot similarity, cosine score, and match likelihood. Molecular networking results were

exported to Cytoscape v3.7.1[103] to visualize and analyze networks. Predicted ClassyFire[104] classifications for shared metabolites were derived using the MolNetEnhancer[105] workflow in GNPS. In addition, select annotations were confirmed using authentic standards (Figure S2.4).

MS filtering was performed in MZmine[100]. Three separate filtering workflows were done: 6 minimum peaks in a row (half the number of samples in a single population), 45 minimum peaks in a row (half our total samples), and 90 minimum peaks in a row (all samples). After each filtering step, gap-filling was performed using the previous parameters. For the six-sample filtering, additional processing was done in R[101] to remove any features that were not found in at least six samples from each population. The resulting files were also analyzed in GNPS as described above.

Mass Spectrometry Search Tool (MASST[106]) was used to search for dataset matches to the MS2 spectra of our shared metabolites. MASST parameters were as follows: parent mass tolerance: 0.02 Da; minimum matched peaks: 4; ion tolerance: 0.5 Da; score threshold: 0.7; top hits per spectrum: 1; selected databases to search: all (GNPS[45], Metabolomics Workbench[107], Metabolights[108], Foodomics[109], and Skin Trace Evidence); no analog searches; and no unclustered data search.

For 16S rRNA gene sequencing data, we used AdapterRemoval v2[110] to filter out sequences < 90 bp in length. QIIME1[111] was used generate ASVs/zero-operational taxonomic unis (zOTUs) using the EzTaxon database[112] for assigning taxonomic identifiers. EzTaxon was selected over other databases like GreenGenes[113] because EzTaxon is regularly updated and taxonomic identification was not the purpose for utilizing our 16S data. All samples with fewer than 10,000 reads were removed from analyses. Any ASVs with detected in fewer than ten samples with a maximum abundance <0.01% were also removed. Generated taxa summaries

were limited to genus-level identifications. Only ASVs with >0.5% relative frequency were included in mmvec analyses.

Mmvec and Statistical Analyses

Metabolite and microbe feature tables were input to the Quantitative Insights Into Microbial Ecology 2 (QIIME2)[114] microbe-metabolite vectors (mmvec) plugin[115]. Conditional probabilities were exported to R[101]. Conditionals were subset to our 67 annotated shared metabolites while major taxa were filtered by exploring high conditional probability values. Filtered results were exported to a new table as .csv. Principal component analyses were run and visualized using the R[101] package "pca3d" [116]. Further figure modifications were done using Inkscape (https://inkscape.org/) v1.2.

Principal coordinate analysis (PCoA) plots were created using Canberra distance metrics from QIIME2[114] and visualized using Emperor[117]. PERMANOVA[118] via QIIME2 assessed statistical significance for beta diversity measures. Kruskal-Wallis p-values were calculated in R[101] through Jupyter Notebook[102]. Boxplots (Figure 2.1c-h, Figures S2.2-2.3) and principal component analyses (Figure 2.3) were also generated using R[101] in JupyterNotebook[102]. For these boxplots, the center line represents the median, the upper and lower box lines reflect upper and lower quartiles, whiskers reflect the interquartile range multiplied by one-and-a-half, and outliers are dots. R packages "ggplot2"[119] and "rworldmap"[120] were used to create Figures 2.1a and 2.1c-h. R package "effectsize"[121] provided p-values for ANOVA effect size. UpSet plots[122] (Figure 2.2b-c; Figure S2.1b) were created using Python 3[123] packages "pandas"[124], "upsetplot"[125], and "matplotlib"[126].

To identify metabolite features unique to specific populations or lifestyles, a random forest machine learning algorithm from the R package "randomForest" was used in Jupyter

Notebook[127]. The number of trees increased gradually from five until reaching a plateau from out-of-bag error at 200 trees. SIRIUS v4.4.26[128] with ClassyFire[104] classification and CANOPUS[129] compound prediction were used to provide class-level annotations for features identified by random forest.

Data Availability

LC-MS/MS data was uploaded to MassIVE[46] (accession number: MSV000084794). GNPS FBMN jobs are available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=505b8b39810c48eb9f9b65fee7c6bc7b (v23, original analysis with gap-filling), https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b76893f1a07e4cb0be3b603c14cea1b2 (v23, gap-filling, primarily used throughout data analysis) and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=af7ec76b02ac482bbd2b7ee3a3ccbdc5 (v23, no gap-filling). FBMN jobs for filtered data are available at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=db26beb51aff418585e6ad0b92f522b7 (six-sample per population filter, gap-filling), https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4693e01a2af740ceb39bfb19720e798d (six-sample per population filter, no gap-filling), https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=220d1afd0a564ec1818601d3d928d27a (half-sample filter, gap-filling), https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d9686d483e5b496299a02750d6a3ec23 (half-sample filter, no gap-filling), https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=45150c751a8e42eea51f3ea4936aee95 (all-sample filter, gap-filling), and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=45150c751a8e42eea51f3ea4936aee95 (all-sample filter, no gap-filling). ReDU co-analysis is available at:

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=cc2c2d20b20d4bd28c22beb777d2782a (co-analysis with all human fecal samples available in ReDU as of August 27th, 2021). This study and associated raw data are available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, Metabolomics Workbench[107] (https://www.metabolomicsworkbench.org; Study ID: ST002320; DataTrack ID 3495; doi: http://dx.doi.org/10.21228/M8N999). MASST search links are provided in Table S2.3. Instructions for recreating data analyses in R and Python are available as JupyterNotebook[102] links at: https://github.com/jhaffner09/core_metabolome_2021. 16S data was uploaded to the Qiita database (study ID: 13802; also see study ID 1442 for Norman, Tunapuco, and Matses data).

2.4 RESULTS AND DISCUSSION

Fecal metabolomes of these populations followed an industrialization gradient, where populations exhibited similar metabolomes based on the degree of industrialization by Principal Coordinate Analysis (PCoA, Figure 2.1b-c, Figure S2.1; Permutational Multivariate Analysis of Variance (PERMANOVA)[118] p=0.001, R^2=0.140; Canberra distance). Moreover, industrialization had a stronger influence on metabolic similarity between populations than geographic origin, age, or sex (Figure 2.1c; ANOVA industrialization group p=0.046, effect size (eta2)=0.06; ANOVA geographic origin p=0.245, eta2=0.01; ANOVA age p=0.4663, effect size (eta2)=7.29e-3; ANOVA sex p=0.5471, effect size (eta2)=4.99e-3). Delay to initial freezing did impact the overall fecal metabolome (PERMANOVA p=0.001, R^2=0.04, ANOVA p=0.4563, effect size (eta2)=6.32e-3), but these effects were overshadowed by the influence of industrialization (Figure 2.1b). For example, the Boulkiemdé rural traditional and Norman urban industrial samples were frozen within one day of collection, but the Boulkiemdé samples

clustered strongly with Peruvian rural traditional samples frozen within four days of collection (Figure 2.1b**)**. Our findings concur with prior studies demonstrating industrialization's role in shaping the human microbiome[95,130–132], the built environment microbiome[43,79], the built environment metabolome[43], and the plasma metabolome[83,133]. This industrialization influence reflects the broad series of behaviors and practices differentiating industrialized, semi-industrialized, and non-industrialized communities such as diet composition and sources, pharmaceutical consumption, hospital access, public markets access, varying exposure to built environments, etc. Further work is needed to determine which components of these industrialized behaviors most contribute to influencing these metabolomes. Nonetheless, the influence of industrialization in differentiating these fecal metabolomes highlights how our molecular biology is strongly affected by our unique lifestyles and behaviors. Additionally, the observation of industrialization outweighing the effects of geographic origin is novel for human fecal metabolomics analyses, but concurs with findings from human fecal microbiome studies[95,130–132]. To the best of our knowledge, this is the first study to illustrate the industrialization gradient in the human fecal metabolome—the intuitive path for revealing the key chemistry of the distal gut.

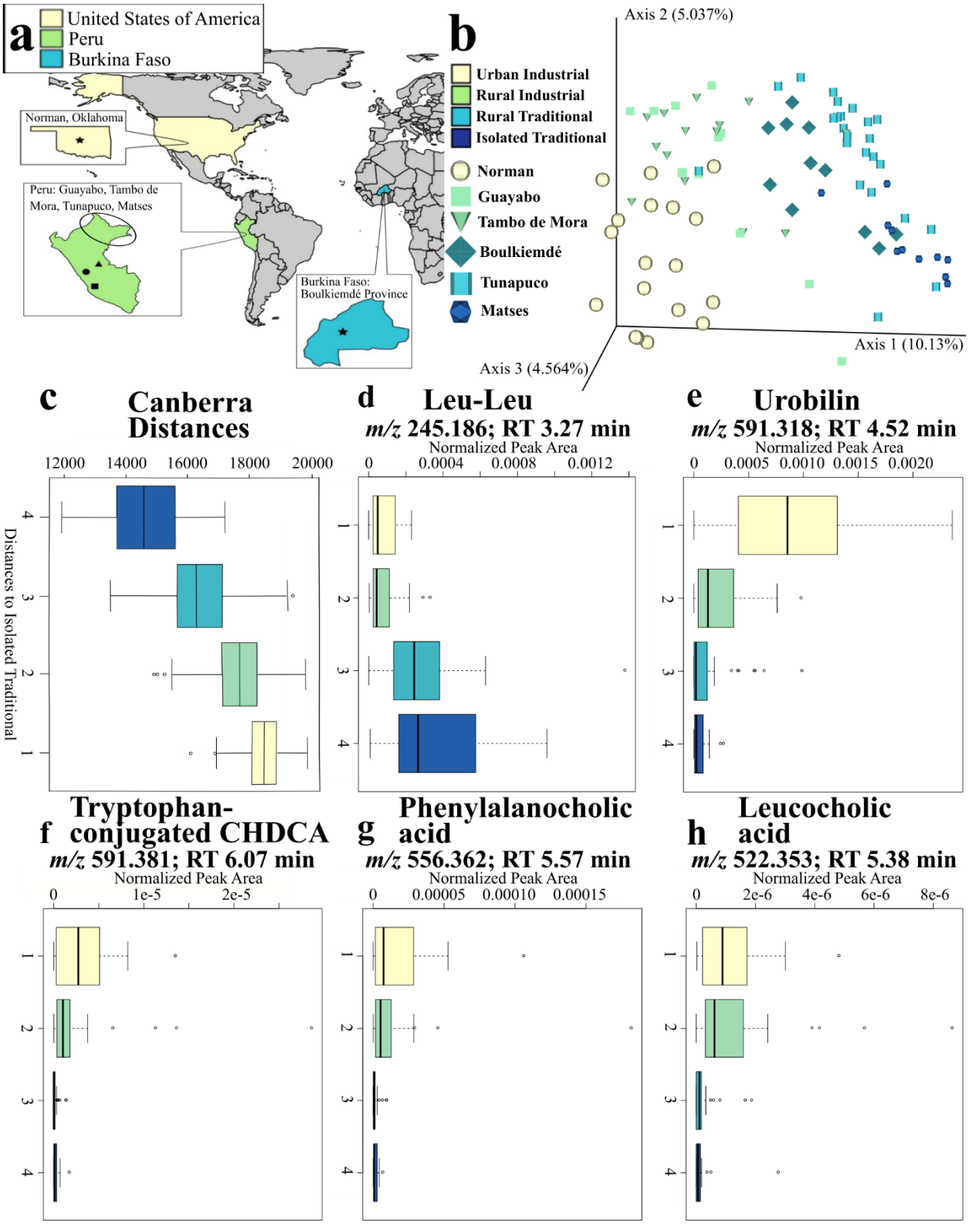

a
United States of America
Peru
Burkina Faso
Norman, Oklahoma
Peru: Guayabo, Tambo de Mora, Tunapuco, Matses
Burkina Faso: Boulkiemdé Province
b
Urban Industrial
Rural Industrial
Rural Traditional
Isolated Traditional
Norman
Guayabo
Tambo de Mora
Boulkiemdé
Tunapuco
Matses
Axis 2 (5.037%)
Axis 1 (10.13%)
Axis 3 (4.564%)
c
Canberra Distances
Distances to Isolated Traditional
d
Leu-Leu
m/z 245.186; RT 3.27 min
Normalized Peak Area
e
Urobilin
m/z 591.318; RT 4.52 min
Normalized Peak Area
f
Tryptophan-conjugated CHDCA
m/z 591.381; RT 6.07 min
Normalized Peak Area
g
Phenylalanocholic acid
m/z 556.362; RT 5.57 min
Normalized Peak Area
h
Leucocholic acid
m/z 522.353; RT 5.38 min
Normalized Peak Area

Figure 2.1: Fecal metabolomic profiles follow an industrialization gradient.
Derived from analyses where sample size (n)=90. ***a****, Sampling sites. Star on tan background: Norman (n=18); Circle on green background: Guayabo (n=12); Square on green background: Tambo de Mora (n=14); Triangle on green background: Tunapuco (n=24); Oval on green background: approximate Matses location (n=11); Star on a blue background: Boulkiemdé (n=11). Specific Matses location left unmarked due to privacy concerns.* ***b****, Principal coordinate analysis (Canberra distance metric) depicts industrialization gradient. Colored by industrialization category and shape-coded by population. Population samples stored in freezer within one day of collection (Norman and Boulkiemdé) are increased in size compared to samples stored within four days of collection (all Peruvian samples—Guayabo, Tambo de Mora, Tunapuco, and Matses).* ***c-h****, boxplot axis numbers represent different industrialization groups: 1=urban industrial, 2=rural industrial, 3=rural traditional, 4=isolated traditional.* ***c****, Calculated Canberra distances follow an industrialization gradient. Colored by industrialization category. Color key from b applies to c-h.* ***d-e****, Normalized abundances of representative features identified by random forest differing by industrialization category:* ***d****, Leucyl-leucine (leu-leu), associated with non-industrialized populations. m/z 245.186; RT 3.27 min.* ***e****, Urobilin, associated with industrialized populations. m/z 591.318; RT 4.16 min.* ***f-h****, Normalized abundances of amino acid-conjugated bile acids depict an industrialization gradient:* ***f****, Tryptophan-conjugated chenodeoxycholic acid (CHDCA). m/z 591.381; RT 6.07 min.* ***g****, Phenylalanocholic acid. m/z 556.36; RT 5.57 min.* ***h****, Leucocholic acid. m/z 522.353; RT 5.38 min.*

To determine the factors driving this clustering of metabolite profiles by industrialization degree, we employed a random forest machine learning algorithm applied to the top 1,000 most abundant metabolite features in our data set[127] (Table 2.2; Figure S2.2a-ad). After applying a variable importance cutoff >1.3 to subset the most differential metabolite features, 377 features remained for annotation. A total of 163 (43.1%) metabolite features had compound-level annotations according to the metabolomics standards initiative[41]. Random forest annotations included glycyl-phenylalanine (mass-to-charge ratio (*m/z)*: 223.108; retention time (RT): 0.38 min; amino acid dipeptide composed of glycine and phenylalanine), piperine (*m/*z: 286.144; RT: 5.45 min; plant metabolite common to pepper plants[134]), and isoleucylproline (*m/*z: 228.155; RT: 0.77 min; amino acid dipeptide detected in human urine[135,136]). When examining the most differential features, two noteworthy annotations were leucyl-leucine (*m/z*: 245.186; RT: 3.27

min; Kruskal-Wallis p=8.73e-09) and urobilin (*m/z*: 591.318; RT: 4.52 min; Kruskal-Wallis p=4.45e-07). Leucyl-leucine (leu-leu) abundance was most associated with non-industrial populations, while urobilin abundance was strongly associated with industrialized populations (Figure 2.1d-e). Leu-leu is a common leucine dipeptide that has not been mentioned in previous industrialization-focused studies of human fecal metabolomes. However, increased abundance of leucine was noted in fecal metabolomes of urban Nigerian adults as compared to rural adults[48], contrasting with the non-industrial association of leu-leu in our data. The second annotated differential metabolite feature, urobilin, is formed from the metabolic breakdown of hemoglobin[137]. While previous industrialization-focused fecal metabolomics studies did not report this metabolite, urobilin has been identified as a common metabolite in human urine and fecal metabolomes[138,139]. Importantly, urobilin abundance is affected by host diet and behavior[140], with increased abundance seen in populations consuming diets rich in animal fat, proteins, and carbohydrates[141], such as those seen in industrialized populations. Given the strong association between industrialization, diet, and the metabolome[80,81,142], it is likely that some unannotated differential metabolite features represent dietary differences between our sampled populations. Meat and processed food consumption was most frequent in industrialized populations, suggesting that any potential dietary metabolites, such as urobilin, likely originate from these industrialized food sources. One such potential dietary source could be artificial sweeteners, which can strongly influence fecal metabolomes[143]. Additionally, the higher consumption of raw vegetable and fruit products in less industrialized communities such as the Matses would also likely drive metabolomic differences. Other potential industrialization-related sources for differential metabolites could include pharmaceuticals and built environment exposure[43,77,79], and gut microbiota modulation of dietary metabolite presence/absence[50,143].

Together, leu-leu and urobilin represent potential biomarkers of an industrialized phenotype. However, the nature of these molecules' association with industrialization remains unclear, especially with regards to health contexts. This association is possibly informed by different dietary behaviors between these industrialization groups, such as urobilin being previously associated with diet[141], but further work is needed to explore the functional role of these molecules in industrialized/non-industrialized contexts. Nonetheless, the altered levels of leu-leu and urobilin based on industrialization highlight metabolomic components of an industrialized phenotype and illustrate potential mechanisms for exploring behavioral effects on human biology.

Recent research has revealed novel amino acid-conjugated bile acids that are produced by gut microbiota[30,106,144] and enriched in patients with inflammatory bowel disease[30]. Given the possible association between inflammatory bowel disease and industrialization processes[88–90], we investigated the distribution of these amino acid-conjugated bile acids across our industrialization gradient. Overall, ten of the 12 total amino acid-conjugated bile acids annotated in this study demonstrated a striking increase with industrialization, despite not appearing in the list of the top 1,000 most abundant features in our data. Such differential amino acid-conjugated bile acids include phenylalanocholic acid (Kruskal-Wallis p=1.9e-6), leucocholic acid (Kruskal-Wallis p=1.69e-7), leucine-conjugated chenodeoxycholic acid (CHDCA) (Kruskal-Wallis p=0.04), tyrosocholic acid (Kruskal-Wallis p=7.71e-3), tyrosine-conjugated deoxycholic acid (Kruskal-Wallis p=1.61e-5), glutamate-conjugated CHDCA (Kruskal-Wallis p=1.69e-7), tryptophan-conjugated CHDCA (Kruskal-Wallis p=4.9e-7), aspartate-conjugated CHDCA (Kruskal-Wallis p=1.13e-5), histidine-conjugated CHDCA (Kruskal-Wallis p=6.41e-3), and histidine-conjugated cholic acid (Kruskal-Wallis p=0.04) (Figure 2.1g-h, Figure S2.2ae-ap).

Interestingly, high abundance of bile acids like phenylalanocholic acid and leucocholic acid were noted in mice fed high-fat diets[30], which is characteristic of Western industrialization societies[145]. The enrichment of these bile acids in our industrialized populations parallels with these diet studies, further suggesting a link between diet and the metabolome across industrialization. Similar to leu-leu and urobilin, these amino acid-conjugated bile acids represent potential biomarkers of an industrialized phenotype, revealing molecular components connected to industrialization. However, two amino acid-conjugated bile acids, aspartate-conjugated cholic acid (Kruskal-Wallis p=0.05) and threonine-conjugated CHDCA (Kruskal-Wallis p=0.4), were not enriched in industrialized populations and did not display any statistically significant differences based on industrialization category. The functional role of these amino acid-conjugated bile acids in health is currently unknown, though our results further support a link between amino acid-conjugated bile acids and industrialization, and possibly to associated diseases.

Feature	*m/z*	RT (min)	p-value (Kruskal-Wallis)	Annotation	Details	Predicted Classyfire/CANOPUS Chemical Class with posterior probability	Mass difference to reference	Adduct	Analog?	Cosine Score
1	145.13	0.321	1.05E-09	-	-	-	-	-	No	-
2	145.13	0.322	1.62E-09	-	Part of same sub-network as Feature 1	-	-	-	No	-
3	159.15	0.359	3.65E-09	-	-	-	-	-	No	-
4	235.17	0.251	2.28E-07	-	-	Primary alcohol (71.332%)	-	-	No	-
5	245.19	3.274	8.73E-09	Spectral Match to Leu-Leu	In sub-network with other Leu-Leu spectral matches	Amino acid derivative (87.591%)	0	M+H	No	0.89
6	276.11	0.411	8.68E-09	-	Part of sub-network with matches to N-Acetylmuramic acid	Organic phosphoric acid and derivatives (59.786%)	-	-	No	-

7	276.11	0.423	1.43E-06	-	Connected to 2 sub-networks, including the sub-network with Feature 6. Also part of network with matches to Glycan Lacto-N-biose and N-Acetylmuramic acid	Organic phosphoric acid and derivatives (59.786%)	-	-	No	-
8	286.18	1.41	4.75E-05	-	-	Secondary carboxylic acid amide (54.113%)	-	-	No	-
9	286.18	1.677	7.10E-06	-	Part of same sub-network as Feature 8	Secondary carboxylic acid amide (54.113%)	-	-	No	-
10	305.19	3.744	2.66E-06	-	-	Carbamate esters (70.111%)	-	-	No	-
11	332.07	0.36	6.36E-08	-	-	Aryl chloride (83.961%)	-	-	No	-

12	363.21	1.018	1.76E-08	-	-	Monosaccharide (59.675%)	-	-	No	-
13	363.21	0.874	1.78E-06	-	Part of same sub-network as Feature 12	Monosaccharide (59.675%)	-	-	No	-
14	365.19	0.514	7.39E-09	-	-	Monosaccharide (56.026%)	-	-	No	-
15	379.3	4.804	4.91E-12	-	-	Lipid and lipid-like molecule (53.344%)	-	-	No	-
16	379.3	4.823	1.17E-10	-	Part of same sub-network as Feature 15	Lipid and lipid-like molecule (53.344%)	-	-	No	-
17	379.3	4.804	4.25E-12	-	Part of same sub-network as Features 15 & 16	Lipid and lipid-like molecule (53.344%)	-	-	No	-
18	379.3	4.811	1.56E-10	-	Part of same sub-network as Features 15, 16, 17	Lipid and lipid-like molecule (53.344%)	-	-	No	-

19	398.34	4.761	1.28E-07	-	-	Fatty acid ester (60.662%)	-	-	No	-
20	398.34	4.829	9.19E-08	-	-	Fatty acid ester (60.662%)	-	-	No	-
21	398.34	4.842	1.15E-07	-	-	Fatty acid ester (60.662%)	-	-	No	-
22	398.34	4.807	1.65E-07	-	-	Fatty acid ester (60.662%)	-	-	No	-
23	400.36	4.832	8.29E-07	-	-	-	-	-	No	-
24	414.34	4.493	3.73E-09	-	-	-	-	-	No	-
25	414.34	4.428	1.18E-09	-	-	Fatty acid ester (63.169%)	-	-	No	-

26	414.34	4.379	1.07E-10	-	-	Fatty acid ester (63.169%)	-	-	No	-
27	414.34	4.428	9.24E-11	-	-	Fatty acid ester (63.169%)	-	-	No	-
28	591.32	4.516	4.45E-07	Spectral match to Urobilin	Part of sub-network with matches to Bilirubin	Fatty acid ester (77.006%)	0	M+H	No	0.79
29	593.33	4.979	3.03E-09	-	-	6-alkylaminopurine (51.054%)	-	-	No	-
30	597.37	5.313	3.27E-06	-	-	Depsipeptide (68.585%)	-	-	No	-

Table 2.2: Top 30 most differential metabolite features as determined by random forest classifier.
These features represent the 30 most differential metabolite features based on mean variable importance scores.

Our sampled populations are considerably different from each other with strong dietary, behavioral, and geographic differences and, together, represent distinct realms of human experience and diversity. Thus, metabolite features common to these markedly separate populations likely constitute shared components of a human fecal metabolome found in major human groups, even if metabolite abundances vary. Frequency assessment of metabolite features can, however, be strongly influenced by data processing parameters, particularly gap-filing and data filtration. Gap-filling identifies peaks that are present in only some samples, and searches for these same peaks at lower intensities in the remaining samples[146]. Analyzing non-gap-filled data can artificially increase divergence between groups, while gap-filling may increase similarities between groups[47,147]. Gap-filling is a recommended approach for feature-based molecular networking[47]. However, to ensure the greatest transparency, we present the analysis of both gap-filled and non-gap-filled data here.

Analysis of non-gap-filled data identified 8,017 metabolite features with at least one occurrence in each population (27,707 common metabolite features in gap-filled data). Further filtering by occurrences in each population highlighted 7,483 metabolite features in non-gap-filled data found in at least six samples in all populations (23,477 metabolite features in gap-filled data), 2,240 metabolite features in non-gap-filled data found in half of all samples in each population (5,924 metabolite features in gap-filled data), and 1,080 metabolite features in both non-gap-filled and gap-filled data found in every sample across all populations (Figure 2.2 for gap-filled data; Figure S2.1 for non-gap-filled data). Impact of industrialization on overall fecal metabolome profiles was comparable between gap-filled and non-gap-filled data (compare Figure 2.1b to Figure S2.1c).

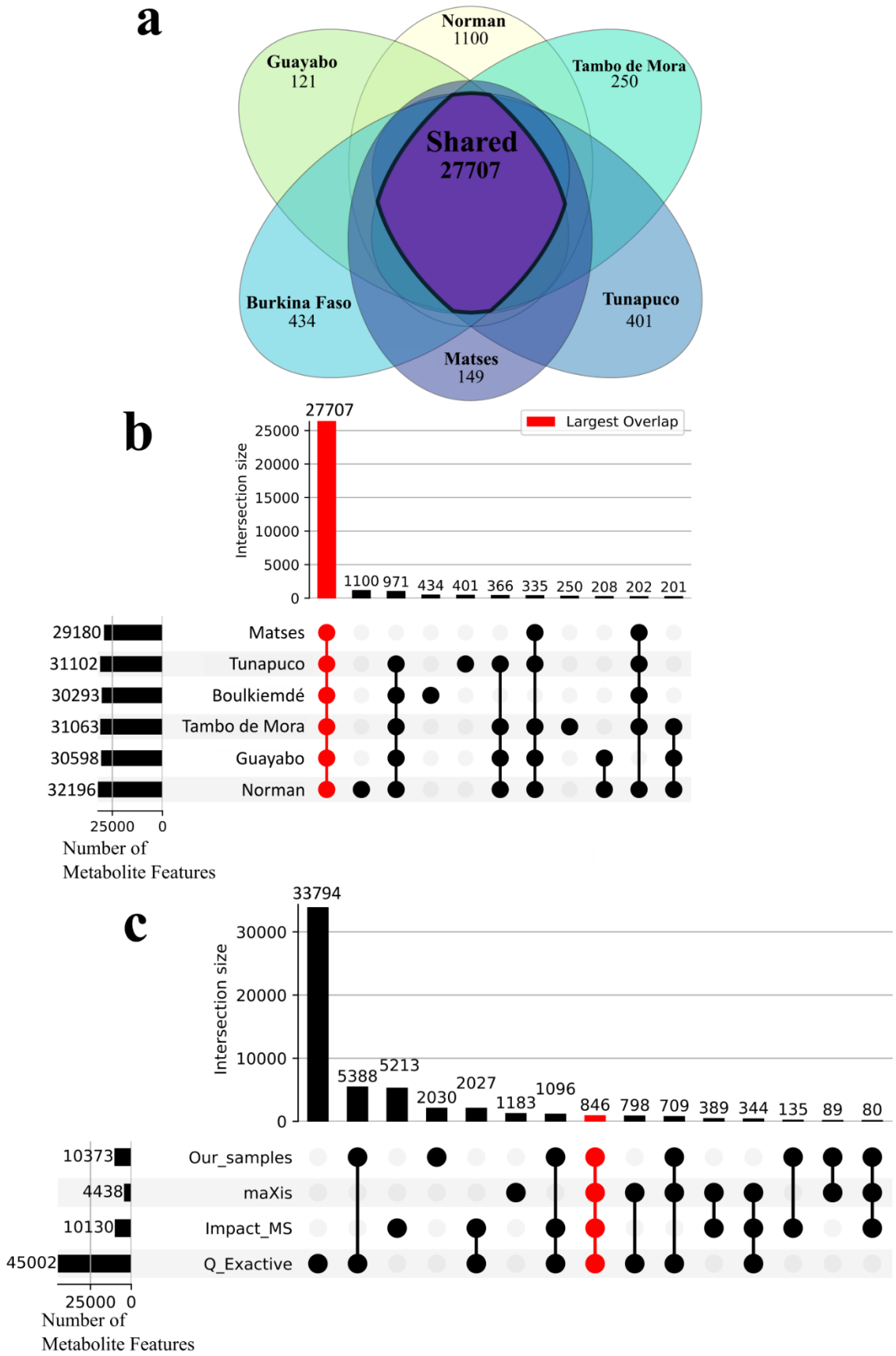

a
Norman
1100
Guayabo
121
Tambo de Mora
250
Shared
27707
Burkina Faso
434
Tunapuco
401
Matses
149
b
Largest Overlap
Intersection size
27707
25000
20000
15000
10000
5000
0
1100 971 434 401 366 335 250 208 202 201
29180
31102
30293
31063
30598
32196
Matses
Tunapuco
Boulkiemdé
Tambo de Mora
Guayabo
Norman
25000 0
Number of
Metabolite Features
c
Intersection size
33794
30000
20000
10000
0
5388 5213 2030 2027 1183 1096 846 798 709 389 344 135 89 80
10373
4438
10130
45002
Our_samples
maXis
Impact_MS
Q_Exactive
25000 0
Number of
Metabolite Features

Figure 2.2: The shared human metabolome.
Derived from where n=90 (Norman n=18, Guayabo n=12, Tambo de Mora n=14, Boulkiemdé n=11, Tunapuco n=24, and Matses n=11). ***a****, Metabolic feature overlap across study populations in gap-filled data.* ***b****, UpSet plot of gap-filled data indicate strong similarity of metabolomic profiles. Total number of metabolite features for each sampled population is depicted in rows with number of overlapping features reported as bar graphs. More features were shared by all populations groups than were seen across different group comparisons. Colored box highlights intersection of all populations (27707 total metabolite features).* ***c****, ReDU co-analysis datasets sorted by MS instrument. ThermoFisher Scientific Q Exactive (n=696); Bruker Impact (n=447); Bruker maXis (n=143). The co-analysis illustrates overlap across the datasets, despite instrumental differences. Colored box highlights intersection of all datasets (846 total metabolite features).*

We discuss here the most stringent results based on the non-gap-filled data, while acknowledging that this approach likely misses many features that are actually common across populations, due to analytical considerations. We further filtered out researcher-derived molecules such as N,N-diethyl-meta-toluamide (DEET) from our list of the shared fecal metabolome. These retained common metabolite features included chemical groups like indoles, steroids, lactones, and fatty acyls (Table S2.2; Figure S2.3). Dipeptides included threonylphenylalanine (*m/z* 267.134; RT 0.48 min), valylvaline (*m/z* 217.155; RT 0.45 min), and isoleucylproline (*m/z* 229.155; RT 0.55 min). Shared bile acids include hyocholic acid (*m/z* 158.154; RT 4.78 min; primary bile acid involved with absorbing and transporting diet fats and drugs to the liver[148]) and lithocholic acid (*m/z* 323.273; RT 6.84 min; secondary bile acid commonly found in feces[149]). Fatty acid examples include 3-hydroxydodecanoic acid (*m/z* 199.169; RT 7.10 min; medium chain fatty acid associated with fatty acid metabolic disorders, potentially acquired from the microbial genera *Pseudomonas, Moraxella,* and *Acinetobacter*[150,151]) and palmitoleic acid (*m/z* 237.001; RT 6.42 min; fatty acid commonly found in human adipose tissue; also acquired in diet from human breast milk[152]). Additional metabolites include cholesterol (*m/z* 369.352; RT 10.5 min; essential sterol found in animals[44]),

methionine (*m/z* 105.058; RT 0.33 min; amino acid), and leucine enkephalin (*m/z* 336.192; RT 3.21 min; peptide naturally produced in animal brains, including humans[44,153]). While a number of these shared metabolite features listed above provide key biological functions, some metabolites appear to be derived from dietary sources. An example of a metabolite possibly acquired from food products includes conjugated linoleic acid (*m/z* 263.24; RT 6.68 min; commonly found in meat and dairy products[44]).

To explore possible interactions between this shared human fecal metabolome and gut microbiome, we used the neural network platform microbe-metabolite vectors (mmvec)[115]. Briefly, mmvec predicts the abundance of metabolites given specific microbial sequences, which then estimates conditional probabilities of co-occurrences between the metabolite and microbe being compared. Given the compositional nature of microbiome and metabolomics data[154,155], mmvec is a robust approach for inferring interactions between gut metabolites and microbes compared to standard correlation analyses[115]. Our mmvec analysis used microbial amplicon sequencing variants (ASVs) derived from earlier sample analyses[96] (see 2.3 Materials and Methods) that were assigned to taxonomic identifications and input to mmvec with our full metabolite feature dataset. After subsetting results to our 67 shared annotated metabolites and their major predictive taxa (27 total), several probable interactions between key gut metabolite features and microbes were observed (Figure 2.3). For example, five microbial species within the *Sporobacter* genus and one unknown member of the Anaeroplasmataceae family were identified as the most influential taxa. Given that five of the six most influential taxa in our dataset were *Sporobacter* species, these results suggest a possible connection between these species and the shared human fecal metabolome. Metabolite features like N-acetyl-L-phenylalanine exhibited strong predictive interactions with an unknown *Sporobacter* species

(Figure 3), as shown by high conditional probabilities. Other strong relationships were observed between abrine and another *Sporobacter* species, as well as glycyltyrosine and N-acetyl-D-mannosamine being strongly driven by the Anaeroplasmataceae member. These potential associations had not been noted in previous literature. All in all, these mmvec results suggest clear patterns of predicted interactions between our shared metabolites and gut microbial taxa, but further work is needed to investigate the connections between the shared human fecal metabolome and gut microbiome, especially with regards to industrialization influence.

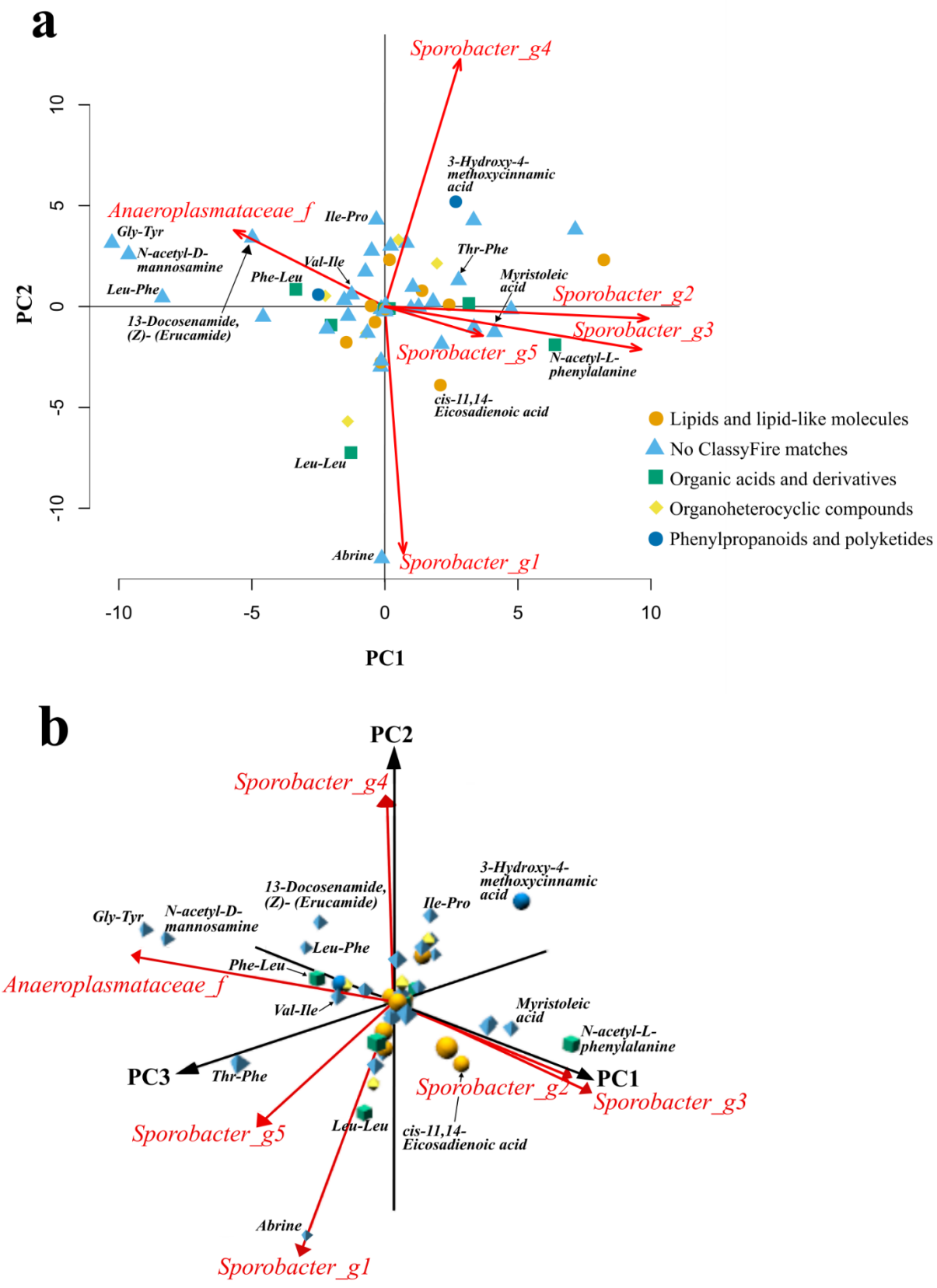

a
Sporobacter_g4
3-Hydroxy-4-methoxycinnamic acid
Anaeroplasmataceae_f
Ile-Pro
Gly-Tyr
N-acetyl-D-mannosamine
Thr-Phe
Val-Ile
Phe-Leu
Myristoleic acid
Leu-Phe
Sporobacter_g2
13-Docosenamide, (Z)- (Erucamide)
Sporobacter_g3
Sporobacter_g5
N-acetyl-L-phenylalanine
cis-11,14-Eicosadienoic acid
Leu-Leu
Abrine
Sporobacter_g1
PC2
PC1
10
5
0
-5
-10
-10
-5
0
5
10
Lipids and lipid-like molecules
No ClassyFire matches
Organic acids and derivatives
Organoheterocyclic compounds
Phenylpropanoids and polyketides
b
PC2
Sporobacter_g4
3-Hydroxy-4-methoxycinnamic acid
13-Docosenamide, (Z)- (Erucamide)
Ile-Pro
Gly-Tyr
N-acetyl-D-mannosamine
Leu-Phe
Phe-Leu
Anaeroplasmataceae_f
Val-Ile
Myristoleic acid
N-acetyl-L-phenylalanine
PC3
Thr-Phe
PC1
Sporobacter_g2
Sporobacter_g3
Sporobacter_g5
Leu-Leu
cis-11,14-Eicosadienoic acid
Abrine
Sporobacter_g1

Figure 2.3: Principal component analyses (PCA) illustrate probable metabolite-microbe co-occurrences.
Derived from analyses where n=90 (Norman n=18, Guayabo n=12, Tambo de Mora n=14, Boulkiemdé n=11, Tunapuco n=24, and Matses n=11). Metabolite feature placements are based on conditional probabilities produced by mmvec[115]. Annotated shared metabolite features from gap-filled data are represented as dots and are shape- and color-coded based on ClassyFire[104] assignments from MolNetEnhancer analyses[105]. PCAs include biplots highlighting the most influential taxa across each principal component (PC) represented with red arrows showing their influence along the PCs. Taxonomic assignments were simplified to include unique identifiers for each label, such as "Anaeroplasmataceae_f" representing a read assigned to the family Anaeroplasmataceae. Multiple Sporobacter genera were identified and were given a "_g" label followed by a number for each instance of Sporobacter genera. ***a****, Two-dimensional representation of shared metabolite-microbe predicted interactions along PCs 1-2. Three taxa are represented for each component. Legend from a applies to b.* ***b****, Three-dimensional figure of shared metabolite-microbe predicted interactions across PCs 1-3. Two taxa are represented for each component.*

Our novel data estimate a shared human fecal metabolome from populations of diverse behaviors and lifestyles. The sample set includes the birthplace of humanity and the last continental expansion of our species, Africa and Americas, respectively; moreover, the sample set includes hunter-gatherer, subsistence farmer, and industrialized lifeways. A metabolite observed across these geographic regions and among these different lifeways is an estimate of a shared metabolome, without implying that it is present in every single individual, similar to definitions used when describing the microbiome[156]. Yet, we do not presume to have captured the complete range of diversity of industrial lifestyles or age groups. To broaden our analysis, we co-analyzed our data with all the publicly-available human fecal samples in the Re-Analysis of Data User Interface (ReDU)[139]. A total of 5,466 human fecal samples from ReDU were co-analyzed with our 90 samples, resulting in a total 105,707 metabolite features detected across the co-analysis (Figure 2c). These datasets contained samples from male and female children and adults. Moreover, the datasets included different MS platforms and different metabolite extraction methods, enabling us to assess the commonality of these metabolites across

experimental methods. Within these datasets, 80% of our shared metabolites were identified in this co-analysis. While these ReDU samples are from datasets containing various instrumental and experimental parameters in addition to wide variation of human diversity and lifestyles, the representation of shared metabolites further highlights the prevalence of a shared human fecal metabolome across different human populations. Furthermore, we also examined the human fecal metabolome database (HFMDB)[42], which contains 6,810 metabolites identified across multiple datasets, for our annotated shared metabolite features. 65% of our annotated shared metabolite features were present in the HFMDB (Table S2.2); examples of identified metabolites also found in the HFMDB include palmitoleic acid, hypoxanthine, and xanthosine. However, it should be noted that the HFMDB is comprised of data derived from various instrumental, analytical, and processing methods[42]. The absence of some of our shared metabolites from the HFMDB can be attributed to these methodological differences.

Furthermore, we used Mass Spectrometry Search Tool (MASST)[106] to search for MS/MS spectra matches to our shared metabolites in public datasets in GNPS[45], Metabolomics Workbench[107], Metabolights[108], Foodomics[109], and Skin Trace Evidence (Table S2.3). These searches report sample types matched with our spectra, such as human, mouse, plant, bacterial, or environmental sample types, as well as matched dataset name. Across our 67 shared metabolites, MASST reported a total of 4,485 total dataset matches, with an average of 67 total dataset matchers per metabolite. Indeed, 61% of our shared metabolites reported more matches to human samples than to other sample types (Table S2.3) and 79% of our shared metabolites also contained bacterial sample matches, suggesting a possible microbiome origin. Additionally, 39% of our shared metabolites were present in human urbanization gradient studies and 67% were present in studies with cultured bacteria from the gut microbiome. Similar to HFMDB and

ReDU, MASST searches contain data collected through various instrumental and experimental methods – any absent shared metabolites can be attributed to these differences. Nonetheless, the MASST searches demonstrate the prevalence of our shared metabolites across different databases and MS/MS platforms.

While we were able to reveal a shared human fecal metabolome, only 6.1% of our complete dataset had putative compound-level annotations (level 2 according to the metabolomics standards initiative[41]). Fifteen of these were validated using standards, enabling level 1 annotation confidence[41] (Figure S2.4). 28.8% of the dataset had annotations based only on chemical class (level 3 of the metabolomics standards initiative[41]). This underscores the need for further annotation of human fecal metabolites, especially from human populations traditionally underrepresented in metabolomic databases. Additionally, it is important to note that samples used for this study were collected at different times and subjected to varying preservation treatments and lengths. However, our samples clustered based on industrialization category rather than storage conditions or geographic origin, indicating that any confounding influence from preservation was overshadowed by the effect of industrialization. Moreover, industrialization refers to a suite of features that can influence the metabolome, with diet as a strong candidate[142,143]. Other factors like demography, genetics, and environment also influence the metabolome with diet, highlighting a need to explore the mechanisms of industrialization's effect on the human metabolome. Furthermore, our sampled populations had unequal sex and age distributions, potentially obscuring any effects caused by sex or age on the fecal metabolome. While our results do not indicate statistically significant differences based on age or sex, further research is needed with samples equally representing sex and age distributions. Full

data are freely available on GNPS[45] so they can be of use to other researchers and annotations can continue to expand.

2.5 CONCLUSIONS

Our results demonstrate how industrialization profoundly shapes human fecal metabolic environments regardless of age, sex, or geographic origin. We also highlight strong commonalities in the fecal metabolome across these distinct populations, representing shared features of a human fecal metabolome represented by endogenous and exogenous metabolites. Based on our definition, these shared chemical components are shared by major human groups or populations, but not necessarily found in every human individual or LC-MS analysis, given differences in metabolite extraction or instrumental conditions between studies. Further studies focused on untargeted analyses of a spectrum of industrial and non-industrial populations, including past and present humans, can help elucidate the shared human fecal metabolome's ubiquity, its relationship with the gut microbiome, and how processes such as industrialization drive human evolution.

Chapter 3 - Untargeted Mass Spectrometry-Based Metabolomics Reveals Responses to High Fiber Whole-Food Diets and Microbiome Composition in Mice[3]

3.1 ABSTRACT

Diet shapes gut microbiome composition and function, which influence the host at least in part by producing metabolites. These metabolites operate on intestinal receptors to cause cascading cellular signals to other organ systems, directly shaping host cellular responses to their environment. Examining organ responses to this metabolite network, called the metabolome, represents a direct link between host functions and phenotypes, the environment (especially diet), and health. In this mouse study, we evaluate the metabolomic consequences of multiple high-fiber diets and microbiome communities. Through a multi-omic approach examining host microbiome and metabolome interactions, our results illustrate how diets and microbiome communities contribute to unique cecum and serum metabolite profiles. Germ-free mice (total *n*=86) were fed standard chow diets for two weeks, then colonized with three different human microbiome communities (subA, subB, subC). Mice were then fed one of three high fiber and high fat combination diets, with legume, vegetable, or whole grain supplementation. After 12 weeks on these diets, cecum contents were sampled and blood serum samples were collected. Metabolites were extracted, separated using C8 liquid chromatography, and analyzed using electrospray ionization mass spectrometry in positive mode with data-dependent MS2 acquisition. Data were analyzed using feature-based molecular networking, spectral library search, principal coordinate analysis, and PERMANOVA. Overall, diet and gut communities

[3] Full author list available in Supplementary Material B.

both differentially modulated cecal and serum metabolites. Analyses focused on microbiome-modulated metabolites that were elevated or depleted compared to the germ-free controls. In the cecum, 40% of all metabolite features were microbiome-modulated, with 16% of these metabolites being microbiome community-specific and 19% diet-specific. The whole grain diet resulted in the greatest number of unique metabolites. In contrast, the microbiome community with the greatest number of unique metabolites varied by diet. Examining diet-donor effects together revealed variation in microbiome-modulated metabolites between donors and diets, with the vegetable diet as the most divergent diet and one specific community (subB) as the most divergent community. Examining molecular networks revealed donor-diet effects on hydroxycinnamic acids (e.g., ferulic acid in whole grain diet), bile acids (e.g., glycocholic acid enriched in subC community) and triterpenoids (e.g., glochidone abundance in legume diet). Furthermore, bile acids were often enriched in legume/vegetable diets, and subC/subA microbiome communities. In the serum, 19% of total metabolite features detected were microbiome-modulated. The whole grain diet was again associated with the greatest amount of unique microbiome-modulated metabolites across donors, but further diversity measures revealed no significant differences between donors across diets. Across these sample types, such diet-based patterns reveal how personalized dietary behaviors varyingly and uniquely shape metabolomes. Moreover, cecum and serum metabolomes exhibited different responses to diet-informed gut microbiota activity. This finding displays spatial aspects to investigating host-microbe relationships, where different metabolome systems are variably influenced by gut microbes and their responses to dietary behaviors. Additionally, significantly different metabolites between experimental groups included structural analogs of curcumol (plant natural product) and glochidone (plant metabolite). . Ultimately, these findings accentuate the strong

potential for dietary effects on personalized health through microbiome-metabolite interactions and highlights mass spectrometry-based metabolomic approaches for examining microbiome-dependent personal dietary effects on host biology.

3.2 INTRODUCTION

Mammalian guts contain diverse microbial communities that influence many aspects of their biology, such as energy harvesting, vitamin production, and behavior[157]. The production of metabolites by gut microbes could influence cellular signaling cascades and cause epigenetic changes by interacting with the intestine and other organs, which in turn affects host physiology[158,159]. Numerous metabolites have been linked to the modification of cardiometabolic outcomes[160].

Whole grain, vegetable, and legume are major sources of fermentable fiber and are frequently encouraged as part of a healthy diet[161]. For example, the positive correlation between legume consumption and the lower risk of cardiovascular disease, obesity, type-2 diabetes, and cancer has garnered considerable interest globally[162]. Additionally, earlier research found the anti-obesity properties of whole grain barley may be mediated by the action of dietary fibers, such as β-glucan, and their capacity to reduce lipid accumulation[163]. However, there was high variation in the benefits associated with whole grain and legume consumption among individuals [164–167]. Dietary fiber effects are distinct due to a number of factors, including individual genetic differences and their gut microbiome composition[168,169]. For instance, intake of whole grains lowers interleukin-6 plasma levels to a greater extent in subjects who had higher levels of *Dialister* and lower levels of *Coriobacteriaceae* species[167]. A recent study linked interpersonal variability in the glycemic response to whole-wheat bread consumption with the gut microbiome differences, suggesting that the relative abundance of *Coprobacter fastidiosus* and

Lachnospiraceae bacterium 3_1_46FAA were predictive features of a lower glycemic response to different bread types[170].

Furthermore, food type, quality, and origin shape gut microbiome composition and function, leading to variable host-microbe interactions[171]. Depending on food source, soluble and insoluble fibers are presented in varying amounts, and not all fiber contents are present in the same food categories[157]. For instance, resistant starch, β-glucans, and arabinoxylans are mostly presented in starchy foods such as oats, cereals and beans, whereas pectin is more abundant in fruits and some vegetables, like carrots and broccoli[172]. Physicochemical characteristics of different dietary fibers, including solubility, viscosity and fermentability, vary widely and influence efficiency in gastrointestinal tracts[173]. Besides fiber, whole grains and vegetables are also rich in polyphenols and vitamins, which modulate the risk of metabolic sydrome[174,175]. Some legume protein fractions, such as Bowman-Birk protease inhibitors, lectins, and seed albumin fractions, prevent intestinal inflammation by modulating gut microbiota[176,177]. Dietary fibers present in whole foods might synergize with these components to exert therapeutic potential in inflammatory and metabolic disorders[157].

Human microbiome research has garnered special interest in anthropology over recent years, particularly with regards to microbial responses to diets[178]. However, studying non-human microbial and metabolomic responses offers unique chances to do deep investigations of personalized health responses under controlled, systematic settings. In this spirit, we sought to investigate how gut microbiome variation contributes to the effect of high-fiber foods on metabolic health in mice. Considering the influence of fiber sources and the interaction with other components, germ-free (GF) mice harboring distinct microbiomes were exposed to three high-fat/fiber/cholesterol diets (45% fat, 7% fiber, 1% cholesterol) that contained fiber from

different whole foods (whole grain, vegetable, and legume). We tested the effects of these high-fiber diets on mice and determined the cecal and plasma metabolites accumulated in the different treatment groups. We found the gut microbiome modulates the effects of high-fiber foods on metabolic health in a highly individual manner, which is reflected in metabolomic data. Additionally, strong individual differences were noted between high-fiber foods, gut microbiota, and cecum metabolites.

3.3 MATERIALS AND METHODS

Experiment Design

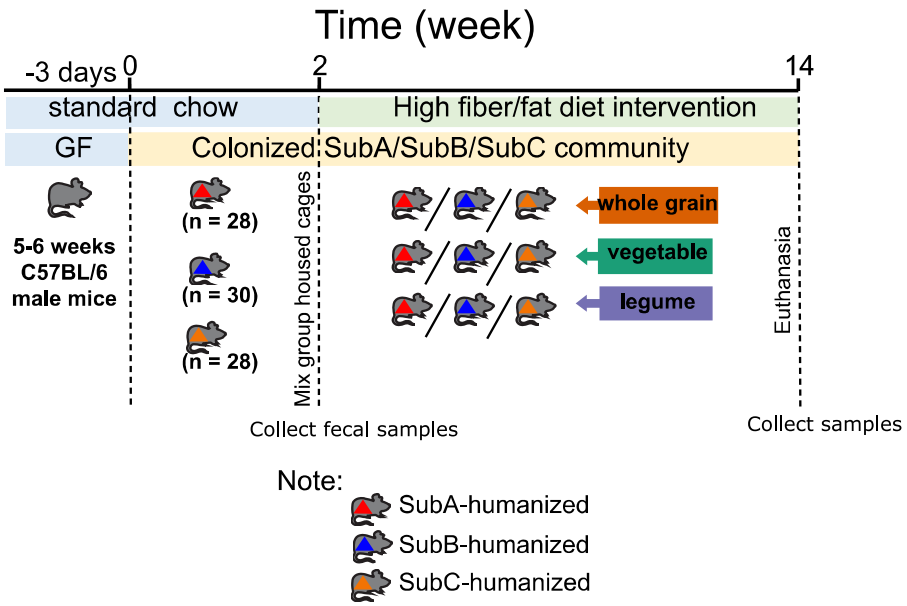


Figure 3.1: Experiment design.

Groups of five- to six-week-old male GF C57BL/6 mice were colonized via oral gavage with one of three human fecal communities: SubA, SubB or SubC (*n* = 28-30 mice/community, n = 86). Human fecal specimens were collected from overweight individuals (BMI = 28-30) aged in their mid-seventies with no history of type II diabetes, cancer, or heart disease and who self-

reported consumption of a standard western-type diet[179]. Selected human communities showed distinct gut microbiome composition and predicted metabolic capacities when engrafted in mice[180] (these human samples were approved for use according to the University of Wisconsin-Madison Institutional Review board, see original publication[180] for full details). For each community, 2-3 batches of animals were colonized at different times. Mice were fed with Teklad global 18% protein rodent diet (chow) a week prior to colonization and for two weeks following exposure to microbes to allow for stabilization. Following this period, mice colonized with each community were divided into three groups (Figure 3.1) and fed one of three isocaloric diets that differed on the type of fiber-rich whole food they contained: (i) whole grains, oats and quinoa trim; (ii) vegetables, carrots and broccoli; or (iii) legumes, black beans and pinto beans (n = 9-10 mice/community/diet). These foods were selected as they represent common sources of dietary fiber in the US[181–183] and were included in the mouse diet as dried powder. Diets provided to the mice also contained a relatively high fat content (45% kcal from diet) and 1% cholesterol to promote metabolic disease[184,185]. Mice were maintained in these experimental diets for 12 weeks. Table 3.1 shows composition for all diets used in the study.

Macronutrients	**Contents**
Crude Protein	18.60%
Fat	6.20%
Carbohydrate	44.20%
Crude Fiber	3.50%
Energy Density	3.1 kcal/g
Calories from Protein	24%
Calories from Fat 18%	18%
Calories from Carbohydrate	58%

Table 3.1: Composition of Teklad global 18% protein rodent diet.

Dietary Formulation

For dietary formulation, we set three high fiber/fat diet groups, including whole grains, vegetables, and legumes. We chose two high fiber foods in each category with the highest consumption in USA. Six food powders bought from Future ceuticals and Van drunen farms were drum-dried and ready to eat. Food powder was then added into three groups of mice pellets as the only fiber source, including broccoli and carrot (vegetables), oats and quinoa (whole grains), black beans and pinto beans (legumes). The three isocaloric diets equally contain 45% Fat, 1% cholesterol, 7% dietary fiber (Table 3.1). Experimental diets were manufactured and sterilized via irradiation by Envigo.

Gnotobiotic husbandry

All experiments involving gnotobiotic mice were performed using protocols approved by the University of Wisconsin-Madison Animal Care and Use Committee. All animal handling was done at the University of Wisconsin-Madison, who conducted mouse experimentation, sampling, and sent treated samples to the McCall Lab at OU. At the University of Wisconsin-Madison, all GF C57BL/6 mice were retained in plastic flexible film gnotobiotic isolators under a twelve-hour light-dark cycle (lights on at 6:00 am and off at 6:00pm) and consuming sterilized water and standard chow (LabDiet 5021; LabDiet) ad libitum until five to six weeks of age. The sterility of GF mice was confirmed by collecting fresh fecal samples and polymerase chain reaction (PCR) gel electrophoresis pattern analysis.

Colonization of GF mice with human fecal samples and diet interventions

Three human communities SubA, SubB and SubC were chosen based on a previous study[180] that showed differences in their gut microbiota diversity. Fecal suspensions of each community were prepared under anaerobic conditions in Hungate tubes. Approximately 500 mg

of frozen feces was resuspended in 5 mL mega media. For each community colonization, ~36 mice at five- to six-weeks of age fed with teklad global 18% protein rodent diet (2918) for 3 days were then orally gavaged with ~100 μL of fecal supernatant. To reduce cage effects, food-filled wires and bedding were switched between cages of mice from the same colonizing community. After two weeks on the 2918 diet, mice were divided into three diet groups and ear tagged. Each group received the indicated diet along with free access to water. Body weights were recorded once a week throughout the experimental period. After 12 weeks, fecal samples were collected. Blood samples were taken through axillary bleeding using isoflurane to induce anesthesia. Serum was acquired by coagulation for 30 min and centrifugation at 1500xg for 10 min at 4 °C and stored at -80 °C until measurement. Prior to sacrifice, mice were fasted for four hours and tissues were collected and snap frozen. Any microbiome analyses and data collection were done by University of Wisconsin-Madison collaborators and were not included as part of this dissertation.

Metabolite extraction for untargeted metabolomics

Metabolites were extracted using an optimized two-step extraction procedure[97,186,187], homogenizing first in 50% methanol spiked with 2 μM sulfachloropyridazine (500 μL/50 mg cecal contents or food samples, or 200 μL/20 μL serum). Homogenate was centrifuged at 16,000xg for 10 min. The centrifugation supernatant was collected and dried overnight in a Speedvac (ThermoFisher). The centrifuged pellet was homogenized in 3:1 dichloromethane: methanol spiked with 3 μM sulfachloropyridazine (500 μL/50 mg cecal contents or food samples, or 200 μL/20 μL serum). Homogenate was centrifuged at 16,000xg for 10 min. The centrifugation supernatant was collected and air-dried. Both dried extracts were stored at -80 °C until LC-MS analysis.

Liquid Chromatography-Tandem Mass Spectrometry Data Acquisition

LC-MS/MS analysis was performed as previously described[187–189]. All solvents used for LC-MS/MS analysis were LC-MS grade, unless otherwise indicated. LC was done using a Kinetex C8 core-shell column (50 x 2.1mm, 1.7 μm particle size, 100 Å pore size) maintained at 40 °C with a ThermoFisher Scientific Vanquish Flex Binary LC system. Samples were kept at 4 °C within the LC system autosampler. For LC, mobile phase solvent A was LC-MS grade water with 0.1% formic acid, and solvent B was LC-MS grade acetonitrile with 0.1% formic acid. A total of 5 μL from each sample was injected in random order. A 7.5-minute runtime elution gradient was used, starting with 2% solvent B for 60 s, gradually increased to 98% solvent B across 90 s, held 98% solvent B for 120 s, decreased to 2% solvent B for 60 s, and held at 2% solvent B for 120 s.

LC system was connected to a ThermoFisher Q Exactive Plus Hybrid Quadrupole-Orbitrap Mass Spectrometer for MS/MS analysis. MS1 and MS2 data were both collected in positive mode. Full MS resolution was set to 70,000, AGC target was 3e6, maximum IT was 246 ms, and scan range was 100-1500 *m/z*. Full MS2 resolution was 17,500, AGC target was 1e5, maximum IT was 54 ms, loop count was 5, and isolation window was 1 *m/z*. Additionally, MS2 data were collected using data-dependent acquisition. The most abundant ion of each MS/MS cycle was identified and 5 MS/MS scans of these ions were recorded. Data-dependent settings were: minimum AGC target of 8e3, intensity threshold of 1.5e5, and dynamic exclusion of 10 s.

Untargeted Metabolomics Data Analysis

Following data acquisition, raw data files were converted to mzXML format using MSConvert[98] v3.019014f9d568a3b. All mzXML files were imported into MZmine[100] v2.53 for detecting sample peaks, building chromatograms, isotope grouping, aligning MS features and

filtering[190]. MZmine parameters were as follows for feature detection: centroided MS1 mass detection, MS1 noise level was 1E5, and MS2 noise level was 1E3; for ADAP chromatogram builder: minimum group size in number of scans was three, group intensity was 1E5, minimum highest intensity was 3E5, *m/z* tolerance was 10 ppm; for chromatogram deconvolution: local minimum parameter enabled, chromatographic threshold was 20%, search minimum in RT range was 0.05 min, minimum relative height was 26%, minimum absolute height was 3E5, minimum ratio of peak top/edge was 1, peak duration range was 0.01-1.1, *m/z* center calculation was set to "median", *m/z* range for MS2 scan pairing was 0.01 Da, and RT range for MS2 scan pairing was 0.1 min; for isotopes peak grouper: isotope peak grouper was 10 ppm, RT tolerance was 0.1 min, charge was 3, and lowest *m/z* was enabled; for alignment: join aligner was 10 ppm, RT threshold was 0.1 min, *m/z* weight was 1, and RT threshold weight was 1; for feature list row filter: keep only with MS2 scan was enabled, RT 0.5-7 min, minimum peaks in a row: 5, and reset ID was enabled.

FBMN was performed using Global Natural Products Social Molecular Networking (GNPS)[45] using the v28.1 GNPS workflow and spectral libraries. For FBMN analysis of MS/MS data, GNPS parameters were as follows: precursor ion mass tolerance was 0.02 Da, fragment ion mass tolerance was 0.02 Da, row sum normalization per file was enabled, mean aggregation for peak abundances per group was enabled, minimum peak filtering intensity was 0, filtered precursor window, filtered peaks in 50 Da window, filtered library, library search minimum matched peaks was 4, cosine score library search threshold was 0.7, maximum analog search mass difference was 100 Da, minimum pairs cosine score for networking was 0.7, maximum number of neighboring nodes was 10, minimum matched fragment ions was 4, maximum connected component size was 100, and maximum shift between precursors was 500 Da (level 2

metabolite annotation per Metabolomics Standards Initiative[41]). Results were exported to Cytoscape v3.9.1[103] for molecular network visualization. The MolNetEnhancer[105] workflow available on GNPS was used after FBMN analysis to improve annotation classifications (level 3 metabolite annotations[41]).

Microbiota-modulated metabolite selection was performed as follows. The resulting feature table was filtered on a per-sample-type basis (cecum or serum) to only retain metabolite features that met one of the following criteria: 1) 100x greater peak area in at least one colonized mouse compared to blanks, food or germ-free mouse samples (microbiota-elevated features) or 2) 100x greater peak area in germ-free mice than in at least one colonized mouse sample AND 100x greater peak area than in blank samples (microbiota-depleted features). The subsets were then combined into a single dataset of microbiota-modulated features, which was used for FBMN[190].

Feature abundance tables were exported into QIIME2[114] qza format. QIIME2 produced Bray-Curtis dissimilarity matrices, which generated principal coordinate analyses (PCoA) plots visualized in EMPeror[117]. Statistical significance comparisons of beta diversity between donor and diet groups were done using PERMANOVA[118] in QIIME2. Venn diagrams were generated by inputting cluster index lists for sample groups at http://bioinformatics.psb.ugent.be/webtools/Venn/. Kruskal-Wallis one-way analysis of variance tests were performed in R, with separate analyses for the cecum and serum datasets. A false discovery rate (FDR)-corrected $p < 0.05$ was used as a significance cutoff. Features differing between specific donor-diet pairs were selected through a median fold change cutoff of ≥ 2. Selected features underwent annotation through GNPS, facilitated by an R plugin (available at: https://github.com/camilgosmanov/GNPS). Annotations were inspected and confirmed/rejected based on mirror plot similarity

comparisons and plausible mass differences based on NetID[191]. Additionally, MolNetEnhancer[105] generated chemical taxonomic assignments from ClassyFire[104] for each metabolite feature. Data were visualized using UpSet plots[122] and colored using chemical taxonomic labels from ClassyFire[104]. Python packages used for making UpSet plots included "pandas"[124], "upsetplot", "matplotlib"[126], "seaborn"[192], and "numPy"[193]. All R and Python code used for Kruskal-Wallis tests, median fold change calculations, file formatting, and UpSet plot generation are available at: https://github.com/jhaffner09/Wisconsin_Diet_Project.

Data availability

Metabolomics data have been deposited in MassIVE, accession number MSV000087045. Molecular networking analysis can be accessed here: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3a3e85a2a63043d7a505547233ea9a91 (full dataset, including food constituents); https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1b1f02ddb16743d091801ca17b83a386 (feature-based molecular networking, microbiota-elevated and microbiota-depleted metabolite features combined); https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=00be111fe8a4480286ba1e1a24dcc814 (MolNetEnhancer). Representative code has been deposited on GitHub, https://github.com/camilgosmanov/GNPS (for automated feature annotation via GNPS) and https://github.com/camilgosmanov/Wisconsin-Diet-Project/blob/main/Diet_Project.R (to filter metabolites depleted in colonized mice).

3.4 RESULTS AND DISCUSSION

Diet and gut community differentially modulate cecal metabolites

We focused specifically on metabolites modulated by the microbiome, indicated as elevation or depletion (>10-fold) compared to GF controls exposed to the same diet (see Methods), representing ~40% of all metabolite features detected in the cecum. We refer to this dataset as microbiome-modulated metabolite features, below. A significant fraction of the microbiome-modulated cecal metabolite features were only detected in a given donor or a given diet (16% and 19%, respectively). The greatest proportion of uniquely elicited metabolites for a given donor was observed in mice fed the whole grain diet (Fisher's exact test, $p < 0.05$ for all pairwise comparisons between whole grain diet and other diets, for a given donor, except for legume *vs.* whole grain in SubB, Figure 3.2E-G).

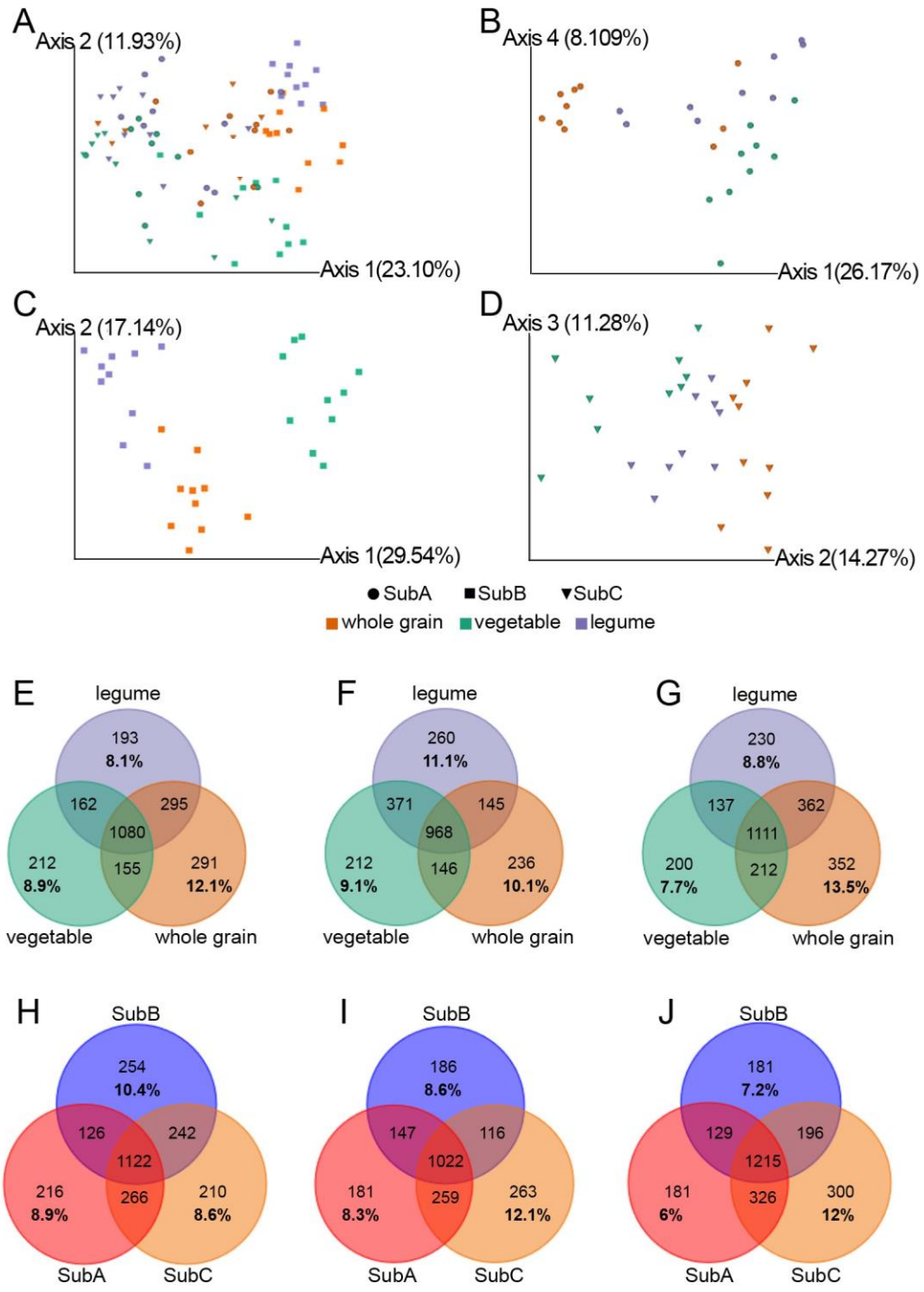


Figure 3.2: Gut microbial community and diet modulate cecal metabolites.
(A) PCoA of Bray-Curtis dissimilarities between cecum microbial metabolites for all three donors. (B-D) PCoA of Bray-Curtis dissimilarities between cecum microbial metabolites for each donor. SubA(B), SubB(C) and SubC(D). Color legend from (A) applies to B-G. (E-G) Numbers of elevated/depleted metabolites present in different diets in each donor. SubA (E),

SubB (F), and SubC (G). (H-J) Numbers of elevated/depleted metabolites present in different donors in each diet. SubA colored red, SubB colored blue, SubC colored orange. Legume (H), vegetable (I), and whole grain (J).

The community with the largest number of unique microbiome-modulated metabolites was variable depending on the diet: SubC for vegetables and whole grain, and SubB for legume diet (Fisher's exact test, $p < 0.05$ for all pairwise comparisons except SubB-SubA for the legume diet, Figure 3.2H-J). Across whole grain and vegetable diets, SubA had the least number of unique microbiome-modulated metabolite features (Fisher's exact test, $p < 0.05$ for all pairwise comparisons except SubA-SubB for vegetable diet, Figure 3.2H-J). The total number of observed microbiome-modulated metabolites (alpha diversity) only differed by diet for SubB, with the lowest number of observed metabolite features in the vegetable diet (Kruskal-Wallis q-value < 0.05 for legume to vegetable and whole grains to vegetable diets, Figure S3.1). Differences between diets were not significant for SubA and SubC, and neither were differences between donors for any given single diet (Kruskal-Wallis q-value > 0.05, FigureS3.1).

In contrast, we observed substantial variation in microbially-modulated metabolite abundance between mice colonized with different communities and diets (Bray-Curtis dissimilarity metric; PERMANOVA p=0.001 R^2=0.196 for diet; p=0.001 R^2=0.174 for donor; p=0.001 R^2=0.0783 for the interaction between diet and donor, Figure 3.2A). The vegetable diet was the most divergent (pairwise PERMANOVA pseudo-F of 10.0 for legumes vs vegetable and 12.0 for whole grains vs vegetable; pseudo-F of 8.4 for legumes vs whole grains; all q-values=0.001). SubB was the most divergent overall, with SubA closer to SubC (pairwise PERMANOVA pseudo-F of 9.4 for SubB vs SubA and of 13.6 for SubB vs SubC, both q-value=0.0015; pseudo-F of 3.2 for SubA vs SubC, q-value=0.003). Within each donor, all diets

differed significantly from each other (pairwise PERMANOVA q-value<0.05), with the most divergent diet varying by donor: whole grain for SubA and SubC, legume for SubB (Figure 3.2B-D).

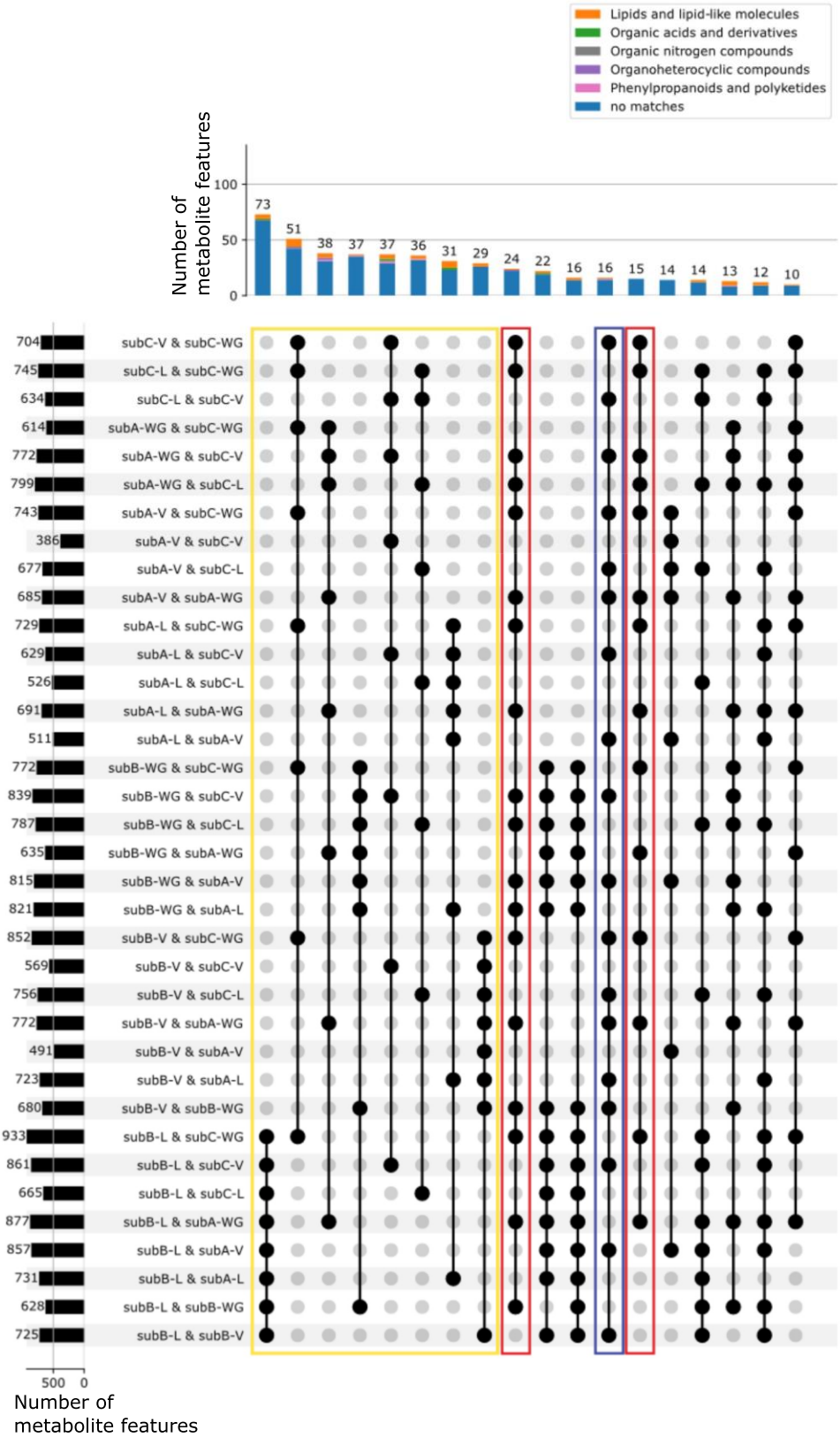

Lipids and lipid-like molecules
Organic acids and derivatives
Organic nitrogen compounds
Organoheterocyclic compounds
Phenylpropanoids and polyketides
no matches
Number of metabolite features
100
50
0
73
51
38
37
37
36
31
29
24
22
16
16
15
14
14
13
12
10
704 subC-V & subC-WG
745 subC-L & subC-WG
634 subC-L & subC-V
614 subA-WG & subC-WG
772 subA-WG & subC-V
799 subA-WG & subC-L
743 subA-V & subC-WG
386 subA-V & subC-V
677 subA-V & subC-L
685 subA-V & subA-WG
729 subA-L & subC-WG
629 subA-L & subC-V
526 subA-L & subC-L
691 subA-L & subA-WG
511 subA-L & subA-V
772 subB-WG & subC-WG
839 subB-WG & subC-V
787 subB-WG & subC-L
635 subB-WG & subA-WG
815 subB-WG & subA-V
821 subB-WG & subA-L
852 subB-V & subC-WG
569 subB-V & subC-V
756 subB-V & subC-L
772 subB-V & subA-WG
491 subB-V & subA-V
723 subB-V & subA-L
680 subB-V & subB-WG
933 subB-L & subC-WG
861 subB-L & subC-V
665 subB-L & subC-L
877 subB-L & subA-WG
857 subB-L & subA-V
731 subB-L & subA-L
628 subB-L & subB-WG
725 subB-L & subB-V
500
0
Number of metabolite features

Figure 3.3: Unique and overlapping metabolite features with differential abundance between donor and diet combinations in cecal contents (UpSet plot).
Kruskal-Wallis, FDR-corrected, $p<0.05$ and fold change ≥ 2 between donor-diet combinations. Connected black circles represent experimental diet-donor pairs that had significantly modulated metabolite features in common (intersections between groups). Figure subset to only include overlaps containing ten or more features. Colored columns represent the number of microbiome-modulated metabolite features in each intersection, colored by ClassyFire[104] superclass. For example, 73 metabolite features differed in abundance by at least two-fold between SubB under the legume diet and all other diet-donor combinations, in a pairwise fashion. Row bars represent the number of metabolite features in a given set, for example 704 metabolite features that met our cutoff criteria when comparing SubC-colonized mice consuming vegetable diet (SubC-V) to SubC-colonized counterparts consuming the whole grain diet (SubC-WG). Features whose differential abundance was characteristic of a given donor-diet combination are marked by a gold box. Features whose differential abundance was characteristic of specific diets are marked by red and blue boxes (red box, characteristic of the whole grain diet; blue box, characteristic of the vegetable diet). V, vegetable; L, legume; WG, whole grain diets.

Consistent with the considerable beta diversity changes observed, 59% of microbiota-modulated metabolite features significantly differed between donor-diet combinations (Kruskal-Wallis $p < 0.05$, FDR-corrected). Differential metabolites were characteristic of a given donor + diet combination (Figure 3.3; gold box, first eight largest intersections), then to a given diet across donors (Figure 3.3, whole grain diet, highlighted in red box and vegetable diet, highlighted in blue box). Most microbially-modulated features could not be annotated, but of those with annotations were mainly lipids (glycerophosphocholines, acylcarnitines, linoleic acid derivatives, bile acids), small peptides, and plant-derived metabolites (triterpenoids, hydroxycinnamic acids) (Figures 3.4-3.6).

With regards to these plant-derived metabolites, hydroxycinnamic acids (Figure 3.4, top left), for example, were observed in cecal contents, mainly in animals consuming whole grain and vegetable diets, as well as in the whole grain, vegetable and legume foods consumed by the animals. Ferulic acid (both adducts combined) significantly differed in abundance between whole grain diets and the other two diets in all donors (except in SubC where a significant

difference was only observed between the whole grain and vegetable diet, Kruskal-Wallis $p < 0.05$, FDR-corrected). Ferulic acid also differed between SubC and the other two donors consuming both a vegetable and a whole grain diet (Kruskal-Wallis $p < 0.05$, FDR-corrected). Triterpene saponins were detected only in the legume diet and in the cecal contents of animals fed a legume diet (Figure 3.4, top right). Triterpenoids likewise were found in whole grain and legume foods, and mainly in the ceca of animals consuming these two diets (Figure 3.4, bottom).

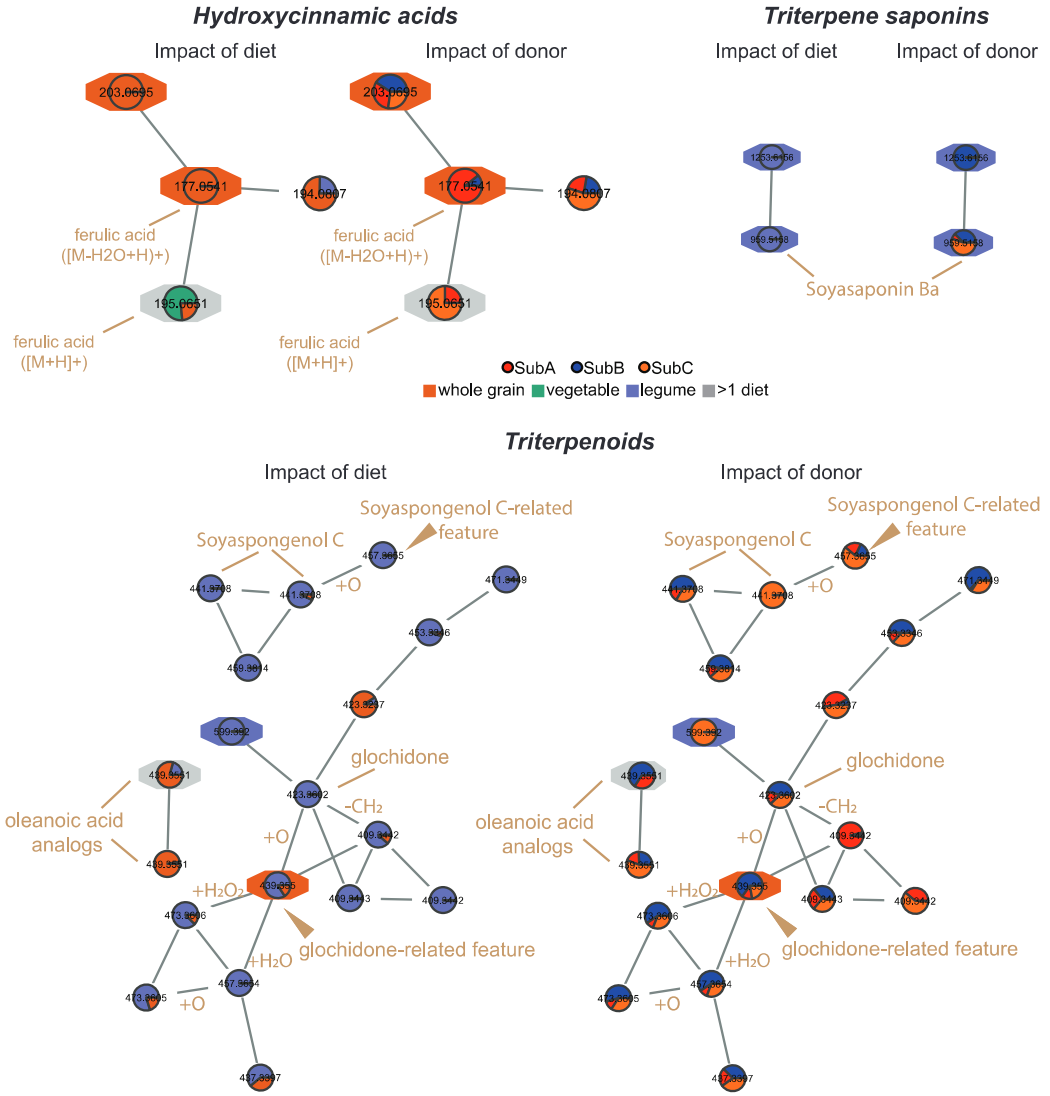


Figure 3.4: Plant-derived metabolite families differentially impacted by microbiome and diet. *Node color background indicates the food in which the node was detected. Clear backgrounds indicate that the node was not detected in food. Node pie charts indicate relative abundance in*

mouse samples, irrespective of sample source (cecum or serum). Black labels indicate node m/z. Node annotations and discussed node chemical mass differences are indicated in brown. Connected nodes indicate MS2 similarity (cosine score >0.7), an indicator of shared chemical substructures. Top left, hydroxycinnamic acids; top right, triterpene saponins; bottom, triterpenoids.

Furthermore, through molecular networking[45,190], we were able to observe differential metabolism of these constituents between donors and diets (Figure 3.4). Molecular networking connects metabolite features with similar MS2 fragmentation patterns, indicating shared chemical substructures, which can indicate metabolism of one metabolite feature into a product metabolite[45,190]. For example, the triterpenoid metabolite *m/z* 439.355 RT 2.78 min (brown arrowhead) was detected in the whole grain food samples and in cecum contents of legume- and whole grain-fed animals. This metabolite feature was directly connected through molecular networking to a deoxygenated metabolite, *m/z* 423.3602 RT 3.05 min, annotated as the triterpenoid glochidone, only detected in cecal contents for legume-fed animals, at comparable levels between donors ($p > 0.05$, Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected, for comparisons between the different legume-fed donors). This could indicate expression of dehydroxylase specifically in legume-fed animals, or a hydroxylase under both legumes and whole grain diets. Dehydroxylation of diet-derived phenols is a known enzymatic activity of the gut microbiome[194]. Multiple other metabolites related to *m/z* 439.355 RT 2.78 min by oxidation/hydroxylation were also detected across donors, in legume and whole grain-fed animals. An oxidized metabolite related to the triterpenoid soyaspongenol C, *m/z* 457.3655 RT 2.88 min (brown arrowhead), was likewise specifically detected in the cecum of legume-fed animals only with no significant difference in abundance between donors. This could indicate expression of an oxidase enzyme specifically in legume-fed animals. In contrast, a demethylated analog of glochidone was observed at significantly higher levels in SubA compared to other

donors (Figure 3.4; $p < 0.05$, Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected, for comparisons between the different legume-fed donors). This may reflect a demethylase enzyme unique to microbes present in the SubA community.

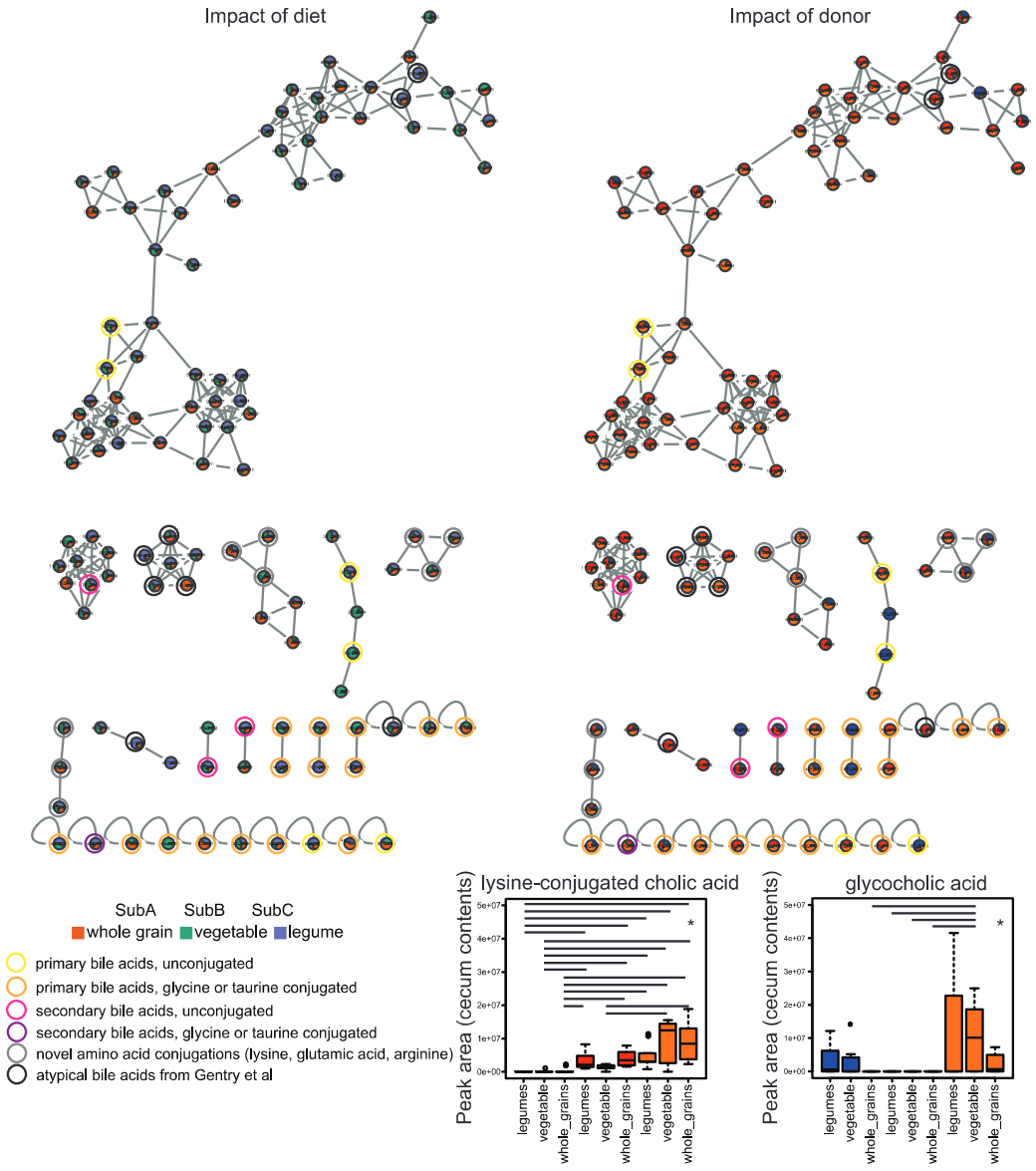


Figure 3.5: Impact of donor and diet on bile acid relative abundance.
*Node pie charts indicate relative abundance in mouse samples, irrespective of sample source (cecum or serum). Labels indicate node m/z. Connected nodes indicate MS2 similarity (cosine score >0.7). *, $p<0.05$ by Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected.*

Other known microbiota metabolites also differed as a function of donor and diet, with considerable variation in primary and secondary bile acid family members. Lysine-conjugated cholic acid, for instance, was significantly lower in SubB compared to SubA and SubC and glycocholic acid was significantly higher in SubC ($p < 0.05$, Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected, Figure 3.5). Of particular interest is commendamide, which is a known G-protein coupled receptor (GPCR) agonist produced by the human microbiome[195] and was elevated across multiple diets in SubA-colonized mice compared to SubB counterparts ($p <0.05$, Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected; Figure 3.6, top).

Kynurenic acid, which is produced under inflammatory conditions and regulates immune responses[196], was significantly higher in SubB-colonized mice consuming legumes and vegetable diets compared to SubC- and to SubA-colonized animals (except for SubA legume diet). Kynurenic acid levels were also higher in SubA-colonized mice consuming the legume diet compared to mice other diets and to mice colonized with SubC community consuming all diets ($p < 0.05$, Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected, Figure 3.6, bottom).

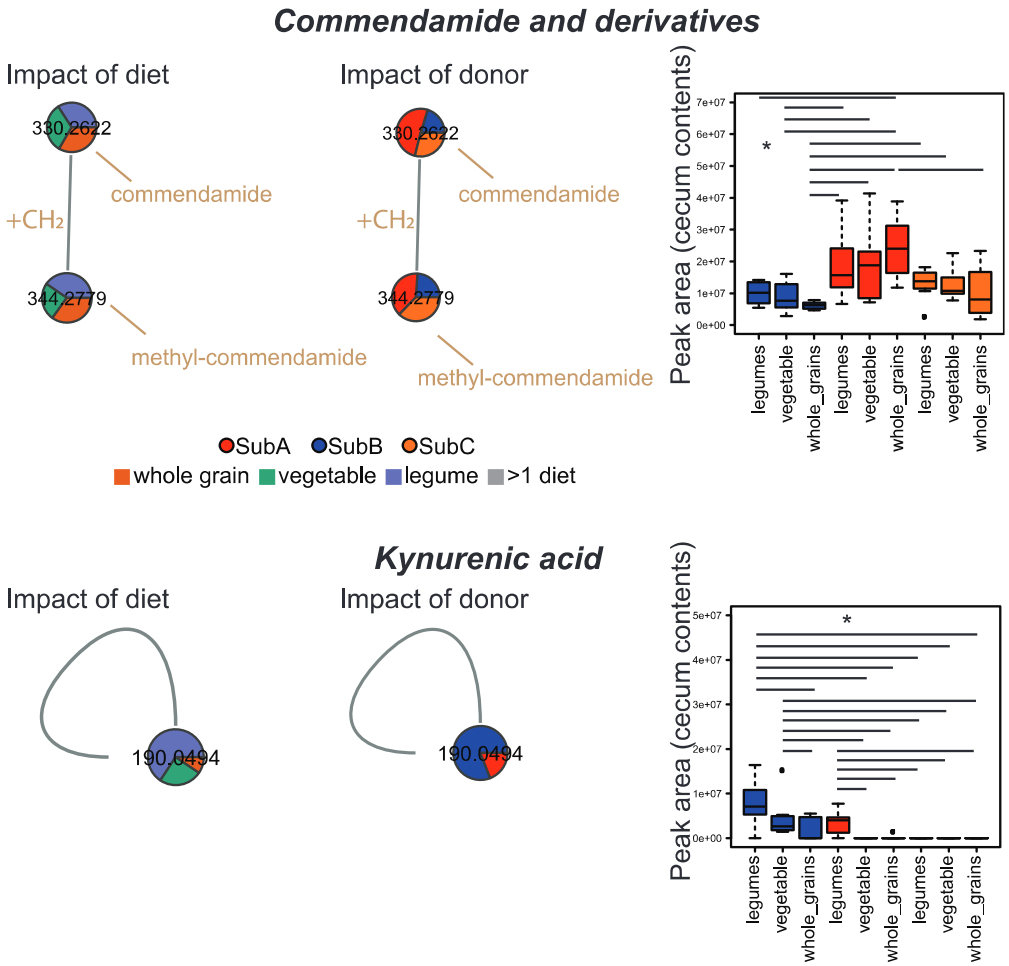


Figure 3.6: Impact of donor and diet on commendamide and kynurenic acid.
*Node pie charts indicate relative abundance in mouse samples, irrespective of sample source (cecum or serum). Labels indicate node m/z. Node annotations and discussed node chemical mass differences are indicated in brown. Connected nodes indicate MS2 similarity (cosine score >0.7). *, p<0.05 by Kruskal-Wallis with post-hoc Dunn's test, FDR-corrected.*

Diets and gut communities modulate serum metabolites

Many of the systemic effects of the gut microbiome are mediated through metabolites that reach the circulation[197,198]. We therefore supplemented the analysis of cecal contents and dietary constituents with metabolomic analysis of serum samples. Serum samples were processed using the same metabolite extraction and data acquisition methods as for food and cecal contents (see Methods). The resulting serum metabolomics dataset was likewise filtered to focus on metabolites modulated by the microbiome, either elevated or depleted compared to serum samples from germ-free controls exposed to the same diet (see Methods). This represented 19% of all metabolite features detected in the serum, a lower proportion than in the cecum. We refer to this dataset as microbiome-modulated metabolite features, and describe their relationship to donor and diet, below.

Overall, the impact of donor community on microbiota-modulated metabolites in the serum was weaker than diet (PERMANOVA p=0.001 R^2=0.129 for diet; p=0.003 R^2=0.052 for donor; p=0.767 R^2=0.034 for the interaction between diet and donor) (Figure S3.2A-D). As with cecal contents, there were far more unique microbiota-modulated metabolites observed under a whole grain diet (Fisher's exact test, $p < 0.05$ for all pairwise comparisons between whole grain diet and other diets for a given donor, Figure S3.2E-G). In contrast, the donors with the least unique metabolite features diverged between diets and did not mirror the observations in the cecum (Figure S3.2H-J). Similar to our observations in the cecum, there were no significant differences in observed metabolite numbers (alpha diversity) between donors, for any given diet (Kruskal-Wallis q-value>0.05). With regards to the impact of diet for a given donor, unlike our observations in the cecum, here we observed greater alpha diversity in the serum of mice receiving whole grain diet for each donor (Kruskal-Wallis q-value<0.05 for comparisons to both

vegetable and legume diets for SubA and SubC; *q*-value<0.05 for comparison to vegetable diet only for SubB; Figure S3.3).

In accordance with the lesser impact of donor and diet on the serum compared to the cecum, fewer microbiota-modulated metabolites significantly differed between donor-diet combinations (13% of serum metabolite features, Kruskal-Wallis $p < 0.05$, FDR-corrected, Figure S3.4). Annotatable differential molecules belonged to the same ClassyFire[104] subclasses as in the cecum, including lipids (bile acids, phospholipids, acylcarnitines) and amino acids, peptides and analogs. Likewise, as in the cecum, many differential metabolites were characteristic of a given donor + diet combination (Figure S3.4, gold boxes). However, only 12 metabolite features met our significance cutoff for both serum and cecum, including glu-val dipeptide (*m/z* 247.1282 RT 0.5 min), a structural analog of curcumol (*m/z* 181.158 RT 3.28 min), and *m/z* 471.3449 RT 2.75 min, structurally related to glochidone (triterpenoid). For most of these overlapping features, however, the impact of diet differed between serum and cecum. Exceptions included: *m/z* 471.3449 RT 2.75 min, which was associated with the legume diet in both sample types; *m/z* 445.2936 RT 2.97 min, which was associated with whole grain diets for both cecum and serum; *m/z* 447.309 RT 3 min, which was also associated with whole grain diet for both sample types; and glu-val dipeptide, which was lower in the whole grain diet compared to the legume diet for SubA and SubC (Kruskal-Wallis with post-hoc Dunn's test, FDR corrected, $p < 0.05$). This indicates the importance of considering local tissue metabolism when analyzing dietary and microbiome composition consequences.

3.5 CONCLUSIONS

Our results demonstrate varying mouse serum and cecum metabolomic consequences of diet-caused microbial responses. Through an emphasis on the molecules modulated by the gut

microbiome, specific diet- and/or microbiome community donor- effects on these metabolomes can be examined. Specific diets, such as the whole grain diet, exhibited strong influences over both metabolomes and resulted in the greatest number of metabolites across donors. Meanwhile, specific donor effects were mixed and were often informed by diet. These investigations of donor-diet effects show the strong influence of individual behaviors, especially diet, in shaping metabolomes. Furthermore, the specific diet-driven metabolomes reveal how diet type, quality, and origin result in different metabolomic phenotypes and illustrate potential pathways for exploring diet-health relationships. Together, these results illustrate specific, individualized responses to high fat and high fiber diets through the gut microbiome, subsequent cecum and serum metabolomic reactions, and how these layers of biology relate together. By conducting this study under controlled settings with mice, the results gleaned from this project represent promising avenues for exploring diet, microbiome, and metabolome relationships in humans. Overall, this project offers a systematic approach for investigating personalized health and metabolomic responses to diets.

Chapter 4 - Mass Spectrometry Database of Archaeologically Relevant Plants for Organic Residue Analysis[4]

4.1 ABSTRACT

Organic residue analysis in archaeology using mass spectrometry (MS) is a robust technique to detect and discuss ancient biomolecules for reconstructing past cultural behavior, such as diet composition and even specific recipes. Studies often involve targeted MS analyses of known or suspected substances, while untargeted analyses characterizing broad ranges of molecules are less common despite their potential to identify unexpected compounds. Regardless of approach, residue studies depend on the quality of reference samples in metabolomics databases, especially with regards to dietary substances like plants, but such reference collections are rarely accessible. This project bridges this gap by generating an untargeted MS-based metabolomics dataset of wild and domesticated plant species from the Americas with significance to archaeological studies, as well as Old World plants. Through generating novel high-quality public plant references from different spatial and temporal contexts, other researchers can compare their unknown samples against our data, increasing molecules identified from plant residues. Moreover, the multi-regional structuring of the plant database allows archaeologists to provide more contextualization to analyses without biasing results to a single region. These data also create opportunities for researchers to explore deeper questions about past and present plant use, regardless of experimental conditions or archaeological context.

[4]Full author list available in Supplementary Material C.

4.2 INTRODUCTION

Archaeological residue analysis is a robust technique for exploring past human cultural behavior. This technique involves any series of biomolecular analyses of organic residues preserved on archaeological samples[21,57], usually focused on identifying chemical compounds within the preserved residue that might indicate the original residue's identity and contents[23,57]. Sample types vary in residue analyses can include bone[199], soil[200,201], and dental calculus[53,202], but ceramics (either intact vessels or fragmented potsherds) are among the most common sample types[58,203–205]. Through determining the chemical makeup of an organic residue, researchers can explore how the archaeological sample was used, interpret potential function of the object (e.g., a fragmented vessel can be interpreted as a smoking pipe if nicotine is detected[62]), identify potential source information, and more[23,57]. While these residue analyses have a long legacy in archaeology, these analyses overlap with techniques and practices seen in MS-based metabolomics[20,62].

In the context of ceramic vessels, these residue analyses often link the presence of detected molecules to the vessel's interpreted use and then assume vessel function, but vessel use does not necessarily equal function[58]. A vessel's use refers to the specific task(s) the vessel was put to, while vessel function refers to the sociocultural-associated contexts of a vessel within the larger archaeological site context[58]. Residue analyses effectively report on vessel use through identifying specific biomolecules found in the preserved residue and apply these detected molecules to interpret vessel use, which is then applied to infer vessel function and study past human behaviors. However, it is important to note that molecular data only reports that the archaeological sample was exposed to that detected molecule at some point in time[23,206]. Associating these detected molecules to vessel use and function, and then interpreting past

human behavior within the larger archaeological context, requires complementing the molecular data with other historical, ethnographic, and/or archaeological data[23,57]. Residue analysis studies have made great strides at navigating the vessel use vs. function dichotomy and exploring past human behavior through these residue data, but these residue-based approaches are nonetheless methodologically limited in a few key areas.

One major limitation with archaeological residue analysis involves the overrepresentation of targeted MS analyses using a biomarker approach[25,28,60–62,207,208]. The biomarker approach in residue analysis treats molecules detected in archaeological organic residues as biomarkers, which are chemical compounds whose presence is assumed to indicate some state (e.g., health or disease states) or past behavior (in archaeological contexts)[21,23]. These targeted residue analyses examine known or suspected molecules[38], treat these detected molecules as biomarkers indicative of past cultural behavior, and then integrate the molecular biomarker data with other existing data to support or reject existing hypotheses[21]. For example, theobromine is a biomarker associated with cacao and chocolate consumption[59,205]. Another common biomarker is nicotine, which is associated with tobacco use[61,62,209]. While these targeted biomarker studies yield useful information about past human behavior, such targeted analyses might miss out on data running counter to existing hypotheses. Untargeted analyses have the potential to detect other interesting molecules, but this analysis type remains less popular in archaeological residue analysis, partly due to lack of existing successful methods.

Another major issue in residue analysis is the reliance on gas chromatography (GC) over liquid chromatography (LC). Beyond the different mobile phase states, GC and LC differ in three crucial ways related to residue analysis: GC uses a heater vaporizer to interact sample molecules with the mobile phase (making GC better-suited for thermally stable molecules), GC is biased

towards nonpolar molecules, and GC is accurate up to 600 Da[15]. Meanwhile, the lack of heater vaporizer for LC makes it ideal for thermally volatile molecules, LC is biased towards polar molecules, and LC is accurate up to 1500 Da. The increased resolution and capacity to detect wider ranges of molecules offered by LC makes it better suited than GC to explore wide ranges of molecules[210,211], especially when applied to untargeted MS analyses. Nonetheless, GC-MS remains the more popular choice for residue analyses[22,23,26,28,212–216]. The popularity of GC-MS over LC-MS has created a methodological gap in residue analysis literature. Although LC-MS is becoming more common in residue analysis[23,59,60,62,208,209,217], the need for LC-based methods still persists, especially for untargeted MS analyses.

Another major challenge with archaeological residue analysis is the lack of standardized high-quality references. In targeted (often GC-MS-based) residue analyses, researchers generally compare the molecule(s) of interest in the data against an authentic chemical standard[23,24,203,205,208,209,215]. Ideally, comparisons against such chemical standards should be done when confirming annotations, according to level one annotations of the MSI[41]. When analyzing samples via MS, data can be compared against MS databases such as GNPS[45] or the National Institute of Standards and Technology (NIST) spectral database[218], which report level two annotations according to the MSI[41]. However, by nature, residue analysis studies are working with degraded organic residues that have been subjected to poorly-understood taphonomic processes[206]. Detected molecules from sampled residues are associated with a specific food product, such as methylxanthine and cacao/chocolate, but few studies incorporate the assumed dietary product (e.g., chocolate) as a reference with the associated organic residue[23,57]. This lack of reference material is especially apparent for residue studies focused on dietary recipes. Some residue approaches adopted experimental archaeology-based techniques, like experimentally-

smoking tobacco as a comparison against historically-smoked tobacco pipes[62] or burying cooking pottery to simulate taphonomic processes on protein residues[203]. While these studies represent steps in the right direction, very few studies analyzed raw, unprocessed plant and/or other dietary material that are associated with residue biomarkers.

In this spirit, we analyzed 281 raw diet samples comprised of wild and domesticated plants from three regions in the Americas. Our goal is to analyze these samples via untargeted LC-MS/MS to create a dataset of high-quality MS references of archaeologically-relevant plant material. Importantly, these data were collected through developing an untargeted metabolite extraction protocol and to prepare samples for untargeted LC-MS/MS; this enables detection of broad metabolite ranges for each sample. While this plant-focused approach is new in archaeological residue analysis, such analyses of dietary products exist in MS-based metabolomics literature. One such study generated chemical inventories of raw and processed foods consumed by modern humans to enable explorations of chemical changes introduced by food preparation processes[39]. However, this study focused on foods common to modern western industrialized human populations[39], while our study examined dietary products relevant to archaeological and Indigenous communities. Through generating untargeted metabolomics profiles of plants and other dietary samples, we expect to see a rise in the number and quality of annotations seen in residue analysis studies. Moreover, detailing the extraction/analysis methods used and making the data accessible create opportunities for other researchers to apply these approaches and data to their own work. While detailed analyses of the plant metabolomic profiles are forthcoming, this dissertation chapter will describe our untargeted LC-MS/MS approach and preliminary results of these plant data.

4.3 MATERIALS AND METHODS

Project Design

A total of 281 plant samples were collected from three locations in the Americas (Figure 4.1): Mexico (*n*=155), New Mexico (*n*=88), and Southern Belize (*n*=34). Sample collection occurred at different times, but sample treatment protocols were standardized across the regions. Plants included grains, fruits, nuts, berries, chiles, roots, inhalants, medicinal and cultural herbs, etc. (see Table S4.1 for full plant list). Additionally, some plants were present across all three regions and multiple parts of plants were subsampled to examine differences based on collected region and different bioactive parts of the plants.

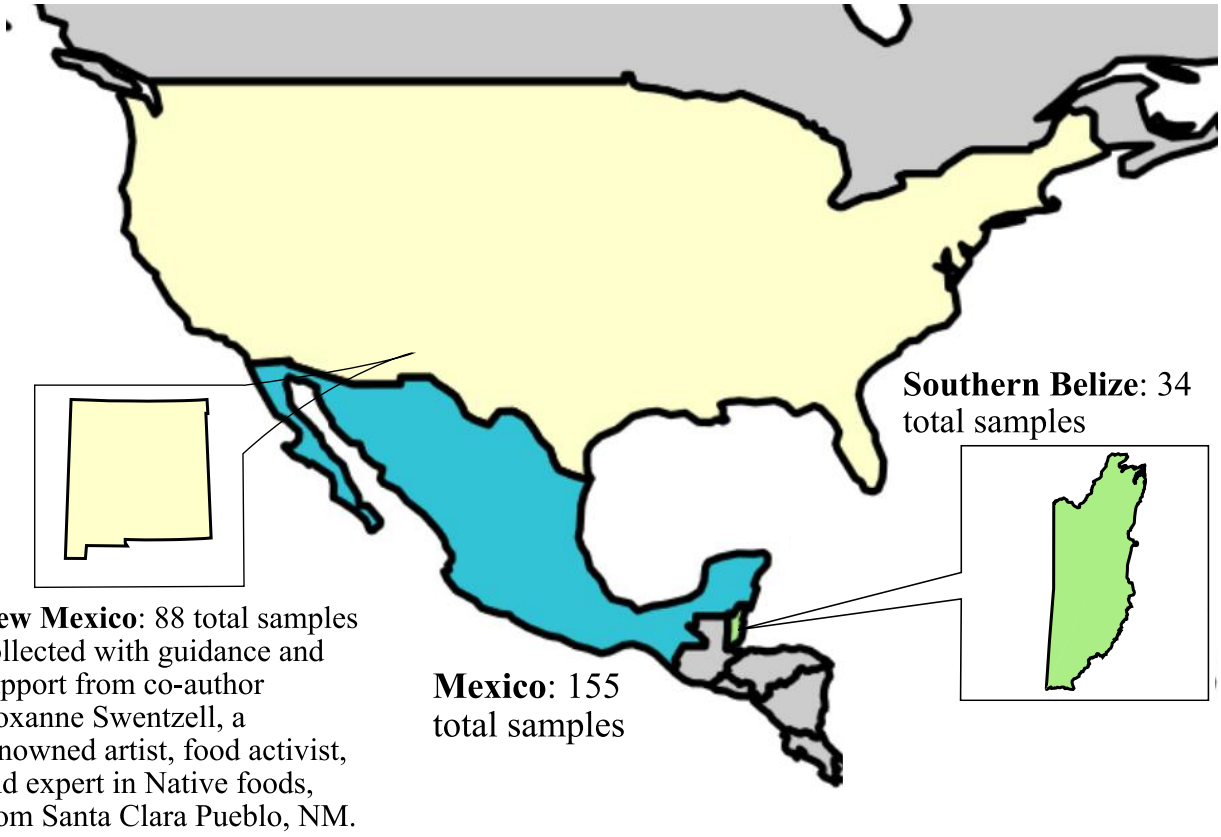


Figure 4.1: Plant sampling locations.

Sample Collection

Mexico samples were collected by co-author Dr. Christina Warinner in 2005-2006 from public markets in two different areas: Mercado Benito Juárez in Oaxaca City (*n*=105, summer and winter collections), and Villahermosa market (*n*=50). All samples were purchased from these public markets. Samples were then subsampled, frozen, and transported by hand to the Harvard University Biochemistry Laboratory. Upon arrival, plants were lyophilized then stored at room temperature. Additionally, samples underwent δ13C and δ15N isotopic analysis by Dr. Warinner[219] prior to work for this project. Samples were then shipped to the Laboratories of Molecular Anthropology and Microbiome Research (LMAMR) in Norman, Oklahoma.

New Mexico plant samples were collected in summer 2022 from two primary locations: the Flowering Tree Permaculture Institute in Española, New Mexico (*n*=83); Plants of the Southwest plant nursery in Albuquerque, New Mexico (*n*=5). Flowering Tree sample collection was directed by Roxanne Swentzell, a renowned artist and expert in local plant history and usage, who provided guidance on which plants were relevant to Indigenous communities historically and today. All Flowering Tree plants were collected on-site. Some plants (namely corns and peppers) were previously sun-dried for storage, while most plants were pulled directly from the ground. Following in-field collection, all non-sun-dried plant materials underwent dehydration using a netted food dehydrator until samples were desiccated. Small aliquots, approximately 50 mg, were stored in 2mL Eppendorf tubes. Following desiccation and subsampling, all remaining plant material were stored in plastic bags. While Flowering Tree samples were drying down, Plants of the Southwest samples were collected from the plant nursery supplies according to direction from nursery employees. The Plants of the Southwest nursery was chosen due to its plant type variety compared to other local nurseries and openness

for sample collection. Plant material was collected directly on-site and stored in plastic bags, according to nursery employee requirements. After sampling, plants were dried down using the same protocol and instrument used with the Flowering Tree samples (all materials were cleaned between usage for the sites). After desiccation, all dried plant material was aliquoted into 2mL Eppendorf tubes. Additionally, four honey samples collected from New Mexico European honeybees (*n*=2) and Yucatán stingless bees (*n*=2) were included to serve as non-plant dietary material. Other non-plant samples included chokecherry-based jam and a soil sample from the Flowering Tree site. Excluding the evaporation step, these non-plant samples were treated identically to the plant samples. After in-field collection and preparation were completed, samples were brought to LMAMR and stored at room temperature.

Belize sample material was collected in 2021-2022 from the Bladen Nature Reserve and the Toledo District of the Santa Cruz Village with guidance from local farmer Jose Mes, who contributed plants from his farm and directed plant identification and collection. After collection, samples were then shipped to the Center for Stable Isotopes at Albuquerque, NM. Upon arrival in Albuquerque, samples were air-desiccated using a netted dehydrator and subsamples were lyophilized for long-term storage. Following desiccation, 50 mg aliquots were stored in 2mL Eppendorf Safe-Lock tubes and brought to LMAMR.

Sample Treatment and Extraction

Once all samples arrived at LMAMR, dried sample aliquots were stored at room temperature. Prior to untargeted metabolite extraction, all subsampled aliquots (~50 mg) were transferred to Sarsedt 2mL screw-cap tubes. The original sample collection was done using 2 mL Eppendorf Safe-Lock tubes, but the Sarsedt 2 mL screw-cap tubes were later preferred because the screw-cap lids better minimize leaks during sample homogenization than the Eppendorf lids.

Extraction protocol was optimized from an untargeted metabolite extraction technique with proven success[97]. One extra step was added: a 24-hour sample soak using extraction solvents. While all samples were analyzed as a single batch via LC-MS/MS, metabolite extractions were done as separate batches based on sample region, with three extraction batches total (one Mexico, one New Mexico, and one Southern Belize).

All used solvents were LC-MS grade. After subsampling, 2 mL of 50% methanol was added to each sample tube. All tubes were then left at room temperature to soak for approximately 24 hours. This 24-hour soak allows the extraction solvent to completely soak and permeate plant material. After the soak, supernatants were removed and stored in separate tubes as backups. The remaining sample material was subject to subsequent treatment for metabolite extraction. Following the supernatant removal, a single 5mm Qiagen steel bead was placed in each sample tube to assist with homogenization. Next, varying amounts of chilled 50% methanol spiked with 2 µM sulfachloropyridazine was added to each tube. Volumes were normalized by sample weight; 50 mg sample corresponds with 1000 µL 50% methanol. After adding extraction solvent, samples were homogenized using a Qiagen Tissuelyzer at 25 Hz for six minutes and centrifuged at 16,000 x g at 4 ºC for ten minutes. Sample tubes were placed on ice while 120 µL aqueous-layer supernatants were collected and placed in a 96-well plate. These aqueous sample supernatants were dried down in a SpeedVac vacuum concentrator to minimize matrix-specific metabolites. All remaining supernatant was transferred to a separate well plate and stored at -80 ºC for long-term storage. The leftover sample pellet (comprised of organic metabolites) was kept as part of the subsequent organic extraction. For organic-layer metabolite extraction, 1000 µL prechilled 3:1 dichloromethane:methanol (spiked with 2 µM sulfachloropyridazine as IS) per 50 mg sample volume was added to sample pellets. Samples were homogenized at 25 Hz for six

minutes and centrifuged at 16,000 x g 4 ºC for ten minutes. After homogenization and centrifugation, 120 µL of each supernatant was placed into a 96-well plate (matching the well position used for the aqueous supernatants) and left to air-dry in a fume hood overnight. The remaining organic-layer supernatant was stored at -80 ºC for long-term storage. After aqueous and organic layers were completely dried down, the sample 96-well plates were stored at -80 ºC until LC-MS/MS analysis. Negative controls with water were included for each extraction batch and underwent identical treatments as plant samples. Sample treatment is also visualized in Figure 4.2.

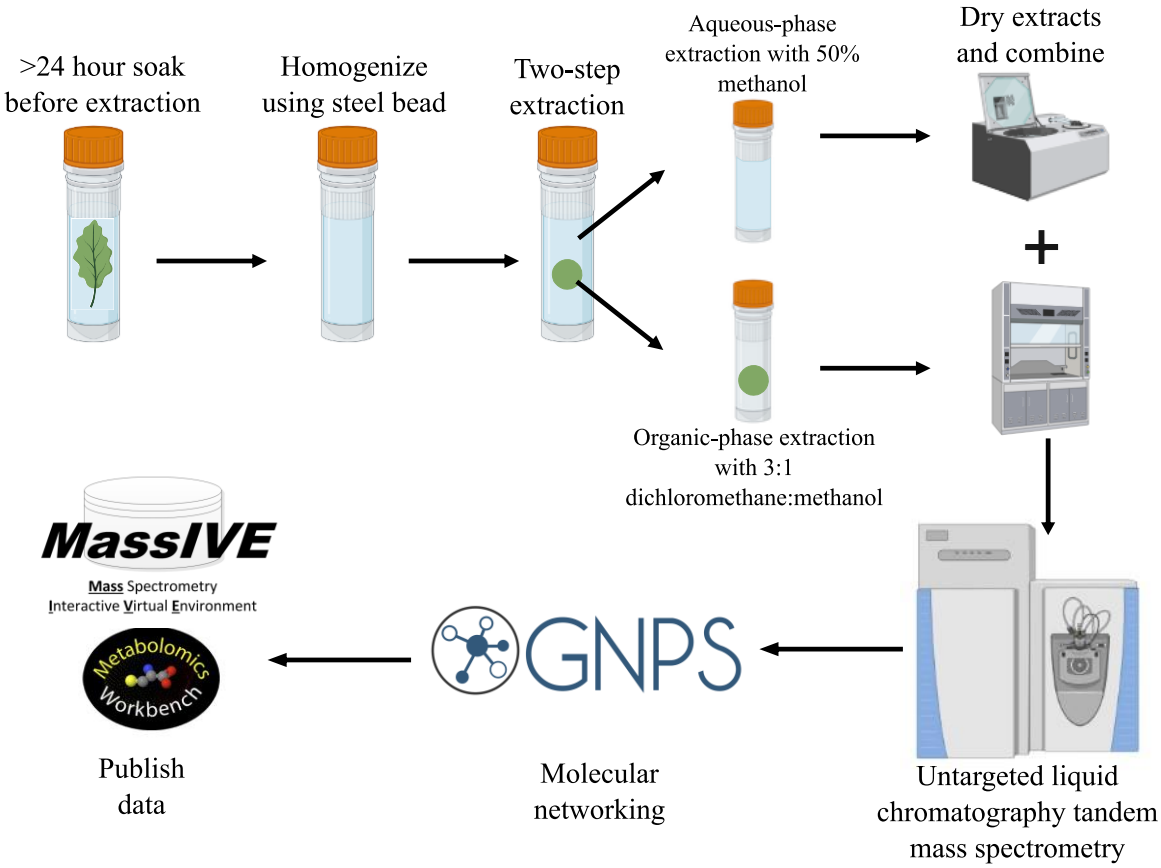


Figure 4.2: Experiment design.

On day of LC-MS/MS run, sample plates thawed at room temperature and 120 µL of 50% methanol spiked with 2 µM sulfadimethoxine (second IS) was added to each well to

resuspend sample material. Sample-matched supernatants were combined and added to a separate 96-well plate for LC-MS/MS analysis. All 96-well plates were sonicated using a Fisher Scientific Ultrasonic Cleaning Bath to dislodge any remaining sample particulates. Each well plate had a single well with 150 µL of 50% methanol to serve as a blank control.

LC-MS/MS Analysis

LC separation was done using a Polar C18 column (Kinetix 100 x 2.1mm, 2.6 µM particle size, 100 Å pore size) kept in a ThermoFisher Scientific Vanquish Flex Binary LC system column compartment maintained at 40 °C. All sample 96-well plates were kept in the LC system's autosampler, which was kept at 4 °C. LC mobile phases were LC-MS grade water with 0.1% formic acid and LC-MS grade acetronitrile with 0.1% formic acid as solvent A and B, respectively. A 12.5-minute runtime gradient was used with the following steps: started at 5% solvent B at 0.5 ml/min and maintained until 1 minute; after one minute, slowly increased solvent B percentage until reaching 100% solvent B by nine minutes total; held at 100% solvent B for two minutes; decreased back to 5% solvent B over 0.5 minutes; and kept at 5% solvent B for the remaining minute. Samples were injected in a randomized order to minimize carryover. Sample injection volume was 5 µL.

MS/MS analysis was conducted using a ThermoFisher Q Exactive Plus Hybrid Quadrupole-Orbitrap Mass Spectrometer. Only positive mode data were collected. Full MS resolution was 70,000, scan range from 70 to 1500 *m/z*, AGC target was 1e6, and maximum IT was 246 ms. For MS2 data analysis, resolution was 17,500, AGC target was 2e6, maximum IT was 54ms, loop count was 5, and isolation window was 1 *m/z*. MS2 data were collected with data-dependent acquisition where the most abundant ion of each MS/MS cycle was identified. For each of these ions, 5 MS/MS scans were recorded. For data-dependent acquisition, minimum

AGC target was 8e3, IT was 1.5e5, and dynamic exclusion was 10 seconds. At start of MS analysis, a well containing resuspension solvent was run and the top 1,000 features were selected to use as an exclusion list where these top 1,000 abundant features were excluded from MS analysis. All subsequent samples were run according to the parameters listed above with the exclusion list.

Data Analysis and Processing

After LC-MS/MS analysis, raw data were converted from .RAW format to .mzXML using MSConvert[98] v v3.0.19014. All files (.raw and .mzXML) were uploaded to MassIVE[46] (also see Data Availability) for storage and to enable classical molecular networking analyses[45]. Molecular networking was done using the "Metabolomics-SNETS-V2" GNPS v30 workflow. Classical molecular networking parameters were as follows: precursor ion mass tolerance was 0.02 Da, fragment ion mass tolerance was 0.02 Da; advanced networking options: minimum pairs cosine score was 0.7, network TopK was 10, maximum connected component size was 100, minimum matched fragment ions was 4, minimum cluster size was 2 with MSCluster enabled, and maximum precursor ion shift was 500 Da; advanced library search options: default spectral libraries were selected, library search minimum matched peaks was 4, analog searches was enabled, minimum cosine score threshold was 0.7, and maximum analog search mass difference was 100; advanced filtering options: filter below standard deviation was 0, filter precursor window was enabled, filtered peaks in 50 Da window as enabled, filtered spectra from Group 6 as blanks before networking was enabled, minimum peak intensity was 0, and filtered libraries was enabled.

Once molecular networking was completed, the MolNetEnhancer[105] v22 workflow was used to estimate chemical taxonomic classifications for level 3 annotations according to the

MSI[41]. MolNetEnhancer data were exported to Cytoscape[103] v3.9.1 to visualize and inspect molecular networks. World map used in Figure 4.1 was made using the "rworldmap" package[120]. Parts of Figure 4.2 (i.e., tubes, MS instrument, etc.) were created with BioRender.com. Statistical analyses were performed using PERMANOVA[118]

Data Availability

Untargeted LC-MS/MS data have been uploaded to MassIVE[46] under the accession number MSV000091215. This dataset is currently set to private and is password-protected, but will be made public once data are ready for publishing with co-author approval. Classical molecular networking analysis is located at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=677b1d18669f43428fb9b8a1e7fd126e (full dataset). MolNetEnhancer molecular networking job is located at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=cdffbdea152e456e9e06b2cf3de57ec5.

4.4 RESULTS AND DISCUSSION

While the primary goal of this chapter is to detail the methods developed for untargeted LC-MS/MS analysis of plant material and to make these data publicly accessible, early analyses reveal interesting patterns. A total of 36,605 metabolite features were detected across these samples using GNPS molecular networking[45]. PERMANOVA[118] analyses indicate metabolomic differences based on region (p=0.001, pseudo-F=7.37794). Considering these sampled regions contained some of the same plants, these results indicate region-based differences in these untargeted plant metabolomic data. Furthermore, significant metabolomic differences were also observed based on the botanical part of the sampled plant (p=0.001, pseudo-F=2.18998). While these results report statistically significant differences for botanical parts, these effects are less pronounced than regional influences. However, these significant botanical part differences

illustrate spatial considerations to plant chemistry. These botanical part differences likely reflect different bioactive areas and functional chemistry. For example, the petals of a flower will exhibit a different metabolome from the flower's stem, which reflect different molecular components and likely different bioactive capabilities. These spatial aspects of plant chemistry should be considered in residue analysis studies, as different plant parts will be varyingly reflected in archaeological residue data. Together, these results illustrate region and botanical part influences on plant metabolomes and emphasize the necessity of spatial considerations when examining plant-derived archaeological residues.

Networking analyses reported annotations of plant-associated metabolite features in these data based on matches to library references (level two annotation, according to the MSI[41]). Such annotations included caffeine (*m/z* 195.024; RT mean 144.6; methylxanthine alkaloid associated with coffee, tea, and cacao[220] – teas and cacao were present in our samples), capsaicin (*m/z* 306.321; RT mean 352.2; associated with *Capsicum* peppers and responsible for their heat, also associated with some health benefits and used as a food additive[221,222] – various *Capsicum* peppers were represented in our dataset), kaempferol (*m/z* 287.054, RT mean 214.8; flavonoid found in a wide variety of plants like teas, squash, cucumber, berries, medicinal plants, etc.[223] – these plants were present in our samples), amygdalin (*m/z* 475.192; RT mean 136.2; found in many plants, but most commonly associated with apples, peaches, and almonds[224,225]), and pinocembrin (*m/z* 257.08; RT mean 362.4; flavonoid detected in many plant types, such as fruits, spices, flowers, teas, etc.[226] Also detected in honey[227] – fruits, spices, flowers, teas, and honey were all present in the data). These annotations and their high-quality matches to the library references (level two MSI annotation[41]) indicate the untargeted metabolite extraction and LC-MS/MS analysis successfully recovered plant-associated molecules from these samples.

Inspecting molecular networks generated from these annotations also reveal region-based differences in metabolite feature abundances. For example, capsaicin was notably enriched in Mexico samples (Figure 4.3A), whereas catechin (*m/z* 291.086; RT mean 170.4; found in plants, especially teas and cocoa[228,229] – teas and cocoa were represented in our data) was highly abundant in Belize and New Mexico plants in one sub-network and enriched in Belize and Mexico plants in a second sub-network (Figure 4.3. Network analyses also illustrated a heightened caffeine abundance in Southern Belize plants (Figure 4.3C) and enrichment of hordenine (*m/z* 194.118; RT mean 174; found in several plant types, particularly associated with algae and cacti[230] – multiple cacti species were part of Mexico and New Mexico sample sets) in Mexico and New Mexico plants (Figure 4.3D), to name a few. Together, these molecular network-derived results reinforce the efficacy of the extraction and analysis methods. Additionally, examining the sub-networks of these plant-associated molecules reveals unannotated compounds with some chemical and/or structural similarity to the detected compound (level 3 annotation according to the MSI[41]). Further work is needed to dig into these molecular networks, especially with regards to the unannotated compounds detected within. Another approach would be to explore how identical plants collected from different regions might exhibit similar or dissimilar molecular networks, shedding light on molecular adaptations to their specific regions. Nonetheless, these molecular networks reflect novel patterns of plant molecular composition.

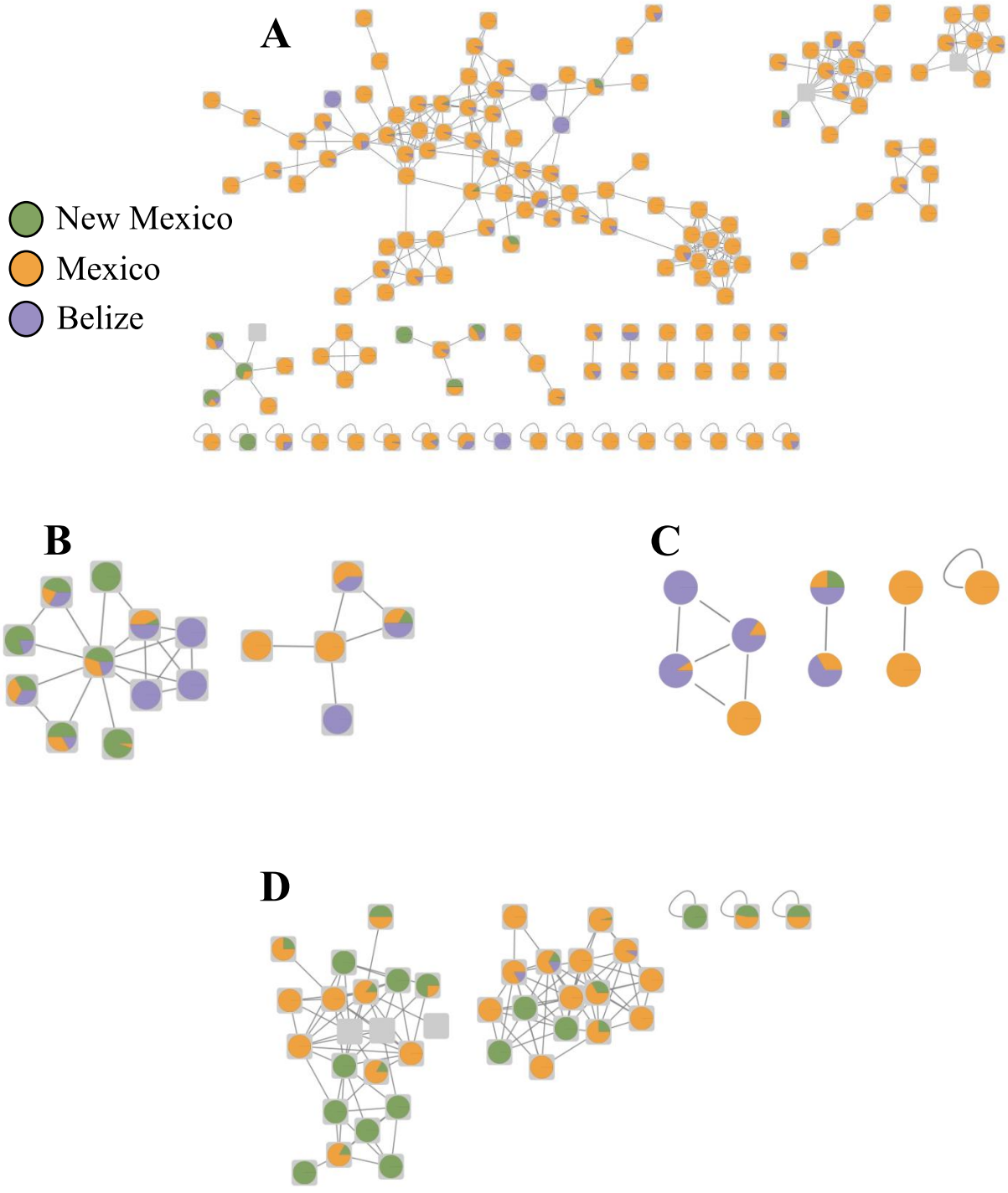


Figure 4.3: Molecular networks highlight region-based abundances of plant chemical compounds.
Nodes represent metabolites structurally related to ***a****, capsaicin ;* ***b****, catechin (m/z 291.086; RT mean 170.4);* ***c****, caffeine (m/z 195.024; RT mean 144.6); and* ***d****, hordenine (m/z 194.118; RT mean 174). Metabolite nodes are depicted as pie charts colored by abundance of that metabolite*

feature in each region – Green: New Mexico; Orange: Mexico; and Blue: Southern Belize. Legend from a applies to b-d. Grey square nodes without pie charts represent metabolite features with unknown identity. Grey lines (edges) reflect similarity between nodes.

While targeted biomarker approaches are commonly used to link degraded archaeological residues to raw plant material, our untargeted metabolome data offers the chance to explore untargeted biomarker-like profiles of plant types as a technique for diagnosing specific plants from ancient residues. For example, detecting plant metabolites unique to cacao varieties can add confidence at identifying the archaeological residue as cacao. Rather than relying on a specific biomarker (or group of biomarkers) from targeted metabolomic analyses, these untargeted profiles contain annotations and unknown molecules with spectral similarity to the known annotations, allowing these plant identifications to be determined by a variety of detected molecules. In this spirit, three untargeted biomarker-like profiles of plant types common to residue analysis literature were created: chiles, maize, and cacao. Unique features from these plant types were identified by inspecting molecular networks and filtering out metabolite features that were uniquely detected in these plant types. Once these unique lists were created, each list was subset to only contain annotations with a mass difference of 0 to the GNPS[45] library reference, indicating a level two annotation according to the MSI[41], and any unknown molecules exhibiting spectral similarity to these annotations, indicating a level three annotation according to the MSI[41]. In total, 18 unique annotations met these criteria in the cacao plants (plus 577 unknown features with spectral similarity), 76 annotations were unique to maize plants (plus 1,322 unknown features with spectral similarity), and 136 annotations were unique to chiles (plus 6,333 unknown features with spectral similarity). Annotations included caffeine (*m/z* 195.023, RT mean 126.96, found in cacao), zeatin riboside (*m/z* 352.161, RT mean 139.72,

found in maize), and quercetin (*m/z* 303.049, RT mean 229.42, found in chiles). Importantly, some of these unique metabolite features represent known biomarkers common to standard targeted biomarker analyses, such as theobromine (*m/z* 361.136, RT mean 56.64) being unique to cacao and capsaicin being unique to chiles. The presence of these known biomarkers within the untargeted biomarker-like profiles demonstrates the utility of these untargeted approaches for helping match molecules preserved in residues to the raw plant material. While further work is needed to fully explore these molecules and their abundances within the plant data, especially through using chemical standards, these untargeted biomarker-like profiles can help match residues to the assumed plant material. These lists will be included when the data are made publicly available.

Our untargeted metabolite extraction method successfully recovered plant compounds expected from these data, but there are opportunities to further optimize the protocol. For instance, one Qiagen steel bead was placed in each sample tube for homogenizing the plant material. However, certain plants, like beans, were tougher to break apart than others. Effective homogenization helps improve metabolite recovery[97] – adding a second steel bead for tougher plant samples could assist the homogenization. Homogenization can also be improved by modifying the Tissuelyzer settings, like increasing the homogenization time length or boosting the shaking frequency. Additionally, all samples underwent a 24-hour soak period with extraction solvents, and this solvent was placed in a separate tube once the 24-hour soak was completed. Rather than removing these 'soaked' supernatants, incorporating them in the metabolite extraction and LC-MS/MS analysis might improve annotation rates since the supernatant likely contains detectable molecules. Furthermore, increasing the amount of time for the soaking period would allow more time for the solvents to permeate sample material and

assist with metabolite extraction, especially if the supernatant is then used for LC-MS/MS analysis. Other potential areas for improving this protocol include a longer runtime gradient to increase LC separation time, modifying extraction and/or resuspension solvents, and grinding up samples during original aliquoting.

4.5 CONCLUSIONS

This dataset represents a series of plant and dietary material relevant to archaeological communities from three different regions of the Americas. By nature, archaeological residues are degraded through taphonomic processes before sampling even occurs and the molecular contents of these degraded residues are associated with raw, unprocessed organic material like plants. However, these residue-plant associations are complicated by a lack of authentic, unprocessed plant material as a reference. This untargeted metabolomic dataset of archaeologically-relevant plants allows direct comparisons between degraded archaeological residues and associated raw, unprocessed plants; it provides wide ranges of plant-derived molecules from known sources as reference samples. By making our data available on popular MS platforms like GNPS[45] and Metabolomics Workbench[107], we want to enable other residue analysis researchers to use these data for comparing against archaeological residues. Ideally, using raw plant/diet material as part of a residue reference database should improve the number and quality of annotations. Publishing these data is our primary goal with this project, and we plan to create panels customized to residue analyses with archaeologists. We also will expand the dataset by including plants from other regions, especially those native to the Africa, Asia, and Europe, as well as processed food materials. This exploration of processed food's chemistry is particularly important considering ancient residues simply report the molecule, or molecules, of interest was exposed to the archaeological sample at some point in time; it is a small component of the

vessel's longer lifespan and exposures to different foods. Extending the data to contain processed foods, food mixtures, and processed food mixtures offers chances to explore the varieties of organic residues and the ways they reflect past human behavior. However, creating untargeted metabolome data of raw, unprocessed plant material represents a necessary first step to deeper investigations of ancient dietary recipes, food preparation/processing techniques, and food chemistry in the past. As residue analysis methods and data improve with the increased adoption of high-quality reference samples, researchers can begin exploring deeper and more complex questions about molecular taphonomy, foodway variations over time and space, and past human behavior.

CHAPTER 5 - CONCLUSIONS

The research discussed in this dissertation represents various ways of applying MS-based metabolomics to address anthropological questions. MS-based metabolomics offers a unique lens into the intersection of human biology with the environment, including how unique lifestyles and their consequences can be gleaned down to a molecular level. With an anthropological context, these chapters highlight specific MS-based metabolomic investigations into aspects of human biology and behavior, both in the present and the past.

Chapter Two, "Untargeted Fecal Metabolomic Analyses across an Industrialization Gradient Reveal Shared Metabolites and Impact of Industrialization on Fecal Microbiome-Metabolome Interactions", provides a novel examination of human fecal metabolomes with a focus on representing variations of human lifestyles, industrialization influence, and geographic origins. By analyzing fecal metabolomes of humans, from western industrialized peoples to traditional hunter-gatherers in the Amazon, we identified metabolomic consequences of different human lifestyles. Specific molecules like urobilin and leucocholic acid illustrate these behavior-biology relationships. Moreover, this research presented new evidence of a shared human fecal metabolome present in major human groups. Incorporating microbial data enabled investigations into how this shared metabolome co-occurs with specific gut microbes like multiple *Sporobacter* species. Together, this chapter detailed the core chemistry of the human gut as it interacts with our unique behaviors while also demonstrating evidence of our shared humanity down to a chemical level.

Chapter Three, "Untargeted Mass Spectrometry-Based Metabolomics Reveals Responses to High Fiber Whole-Food Diets and Microbiome Composition in Mice", took a different approach from traditional anthropological studies by investigating metabolomic mechanisms

associated with diet-induced microbial changes. This project studied how three transplanted microbiome communities responded to legume, vegetable, and whole grain diets by focusing on cecum and serum metabolites modulated by these microbial communities. Specific donor, diet, and donor-diet patterns were gleaned, as evident by the whole grain diet exhibiting the greatest number of unique metabolites across both metabolomes. Meanwhile, the effects of microbiome donor communities on the number of unique metabolite features were more mixed. Unique cecum metabolite features influenced by microbiome donor were varied by diet, but donor effects on the serum metabolome were muted overall. Specific molecule classes like triterpenoids were varyingly influenced by diet and microbiome community, illustrating potential mechanisms for microbial and dietary activity on these metabolomes. Through using mice as a controlled, systematic research environment, these data show new approaches for exploring personalized health responses to diets.

Chapter Four, "Mass Spectrometry Database of Archaeologically Relevant Plants for Organic Residue Analysis", shifted focus to ancient molecules through developing new methods and novel data for archaeological organic residue analysis. In this project, plant samples from different regions significant to archaeologists were analyzed using similar techniques as seen in the previous biology-focused chapters. These plant metabolomes were characterized to provide broad chemical profiles of these plants for archaeologists to reference when doing residue analysis. These data will be made publicly available on a variety of MS databases, but preliminary investigations highlighted common plant-associated molecules, like capsaicin, caffeine, and hordenine, were present in the data. These patterns demonstrate the efficacy of the extraction/analytical methods and reveal differential abundances based on region. Once forthcoming analyses are concluded and the data are publicly accessible, deeper patterns can be

explored, such as molecular expressions of plant adaptations to their region and molecular taphonomy. Overall, the use of these plant data could elevate the number and quality of annotations detected in plant-based archaeological residues.

Taken together, these three chapters present varied research projects using MS-based metabolomics in anthropological contexts. MS-based metabolomics approaches not only provide unique perspectives into the molecular dimensions to human health and biology, but such approaches are uniquely poised to explore the intersection of human health and biology with lifestyle, behavior, and the environment. However, it is important to note that much work remains to integrate MS-based metabolomics approaches with anthropology, especially in ancient contexts. Since molecular taphonomy and preservation is still poorly understood, there is a significant need to develop untargeted MS methods for investigating ancient molecules. Nevertheless, MS-based metabolomics is recognized as increasingly important in biological and health contexts. Understanding the whole of human biology, health, and their connections to lifestyles, both in the past and the present, requires MS-based metabolomics to highlight the molecularly-informed answers to these questions. Ultimately, this dissertation represents the wealth of information that MS-based metabolomics offers to anthropology.

Bibliography

1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omic trilogy. *Nat Rev Mol Cell Biol* **13**, 263–269 (2012).
2. Kottak, P. C. *Anthropology: Appreciating Human Diversity*. (McGraw Hill, 2018).
3. Marks, J. What Is Molecular Anthropology? What Can It Be? *Evol. Anthropol.* **11**, 131–135 (2002).
4. Moeller, A. H., Li, Y., Mpoudi Ngole, E., Ahuka-Mundeke, S., Lonsdorf, E. V., Pusey, A. E., Peeters, M., Hahn, B. H. & Ochman, H. Rapid changes in the gut microbiome during human evolution. *Proc. Natl. Acad. Sci.* **111**, 16431–16435 (2014).
5. Kuzawa, C. W. & Quinn, E. A. Developmental Origins of Adult Function and Health: Evolutionary Hypotheses. *Annu. Rev. Anthropol.* **38**, 131–147 (2009).
6. Millikan, R. The Changing Face of Epidemiology in the Genomics Era. *Epidemiology* **13**, 472–480 (2023).
7. Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Pauvnovic, M. & Pääbo, S. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336 (2006).
8. Ovchinnikov, I. V., Götherström, A., Romanova, G. P., Kharitonov, V. M., Lidén, K. & Goodwin, W. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404**, 490–493 (2000).
9. Hendy, J., Colonese, A. C., Franz, I., Fernandes, R., Fischer, R., Orton, D., Lucquin, A., Spindler, L., Anvari, J., Stroud, E., Biehl, P. F., Speller, C., Boivin, N., Mackie, M., Jersie-Christensen, R. R., Olsen, J. V., Collins, M. J., Craig, O. E. & Rosenstock, E. Ancient proteins from ceramic vessels at Çatalhöyük West reveal the hidden cuisine of early farmers. *Nat. Commun.* **9**, (2018).
10. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–71 (2002).
11. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass Spectrometry-Based Metabolomics. *Mass Spectrom. Rev.* **26**, 51–78 (2007).
12. Viant, M. R., Kurland, I. J., Jones, M. R. & Dunn, W. B. How close are we to complete annotation of metabolomes? *Curr. Opin. Chem. Biol.* **36**, 64–69 (2017).
13. Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S. & Sumner, L. W. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **9**, 418–425 (2004).
14. Wolfender, J., Marti, G., Thomas, A. & Bertrand, S. Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A* **1382**, 136–164 (2015).
15. Greaves, J. & Roboz, J. *Mass Spectrometry for the Novice*. (CRC Press, 2014).
16. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Mol. Cell* **58**, 699–706 (2015).
17. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted metabolomics. *Curr. Protoc. Mol. Biol.* **98**, 1–34 (2012).
18. Matsuda, F. Technical Challenges in Mass Spectrometry-Based Metabolomics. *Mass Spectrom.* **5**, S0052–S0052 (2016).
19. Coskun, O. Separation Techniques: Chromatography. *North. Clin. Istanbul* **3**, 156–160

(2016).
20. Brockbals, L., Habicht, M., Hajdas, I., Galassi, F. M., Rühli, F. J. & Kraemer, T. Untargeted metabolomics-like screening approach for chemical characterization and differentiation of canopic jar and mummy samples from Ancient Egypt using GC-high resolution MS. *Analyst* **143**, 4503–4512 (2018).
21. Evershed, R. P. Organic residue analysis in archaeology: The archaeological biomarker revolution. *Archaeometry* **50**, 895–924 (2008).
22. Pecci, A., Giorgi, G., Salvini, L. & Cau Ontiveros, M. Á. Identifying wine markers in ceramics and plasters using gas chromatography-mass spectrometry. Experimental and archaeological materials. *J. Archaeol. Sci.* **40**, 109–115 (2013).
23. Whelton, H. L., Hammann, S., Cramp, L. J. E., Dunne, J., Roffet-Salque, M. & Evershed, R. P. A call for caution in the analysis of lipids and other small biomolecules from archaeological contexts. *J. Archaeol. Sci.* **132**, (2021).
24. Zagorevski, D. V. & Loughmiller-Newman, J. A. The detection of nicotine in a Late Mayan period flask by gas chromatography and liquid chromatography mass spectrometry methods. *Rapid Commun. Mass Spectrom.* **26**, 403–411 (2012).
25. Steele, V. J. & Stern, B. Red Lustrous Wheelmade ware: Analysis of organic residues in Late Bronze Age trade and storage vessels from the eastern Mediterranean. *J. Archaeol. Sci. Reports* **16**, 641–657 (2017).
26. Giorgi, G., Salvini, L. & Pecci, A. The meals in a Tuscan building yard during the Middle Age. Characterization of organic residues in ceramic potsherds. *J. Archaeol. Sci.* **37**, 1453–1457 (2010).
27. Salvini, L., Pecci, A. & Giorgi, G. Cooking activities during the Middle Ages : organic residues in ceramic vessels from the Sant ' Antimo Church (Piombinoo-Central Italy). *J. Mass Spectrom.* **43**, 108–115 (2008).
28. Charters, S., Evershed, R. P., Goad, L. J. & Leyden, A. Quantification and Distribution of Lipid in Archaeological Ceramics: Implications for Sampling Potsherds for Organic Residue Analysis and the Classification of Vessel Use. *Archaeometry* **35**, 211–223 (1993).
29. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
30. Quinn, R. A., Melnik, A. V., Vrbanac, A., Fu, T., Patras, K. A., Christy, M. P., Bodai, Z., Belda-Ferre, P., Tripathi, A., Chung, L. K., Downes, M., Welch, R. D., Quinn, M., Humphrey, G., Panitchpakdi, M., Weldon, K. C., Aksenov, A., da Silva, R., Avila-Pacheco, J., Clish, C., Bae, S., Mallick, H., Franzosa, E. A., Lloyd-Price, J., Bussell, R., Thron, T., Nelson, A. T., Wang, M., Leszczynski, E., Vargas, F., Gauglitz, J. M., Meehan, M. J., Gentry, E., Arthur, T. D., Komor, A. C., Poulsen, O., Boland, B. S., Chang, J. T., Sandborn, W. J., Lim, M., Garg, N., Lumeng, J. C., Xavier, R. J., Kazmierczak, B. I., Jain, R., Egan, M., Rhee, K. E., Ferguson, D., Raffatellu, M., Vlamakis, H., Haddad, G. G., Siegel, D., Huttenhower, C., Mazmanian, S. K., Evans, R. M., Nizet, V., Knight, R. & Dorrestein, P. C. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).
31. Tan, K. C., Pham, T. X., Lee, Y., Lee, J. & Balunas, M. J. Identification of Apocarotenoids as Chemical Markers of In Vitro Anti-In fl ammatory Activity for Spirulina Supplements. *J. Agric. Food Chem.* **69**, 12674–12685 (2021).
32. Petras, D., Nothias, L. F., Quinn, R. A., Alexandrov, T., Bandeira, N., Bouslimani, A., Castro-Falcón, G., Chen, L., Dang, T., Floros, D. J., Hook, V., Garg, N., Hoffner, N.,

Jiang, Y., Kapono, C. A., Koester, I., Knight, R., Leber, C. A., Ling, T. J., Luzzatto-Knaan, T., McCall, L. I., McGrath, A. P., Meehan, M. J., Merritt, J. K., Mills, R. H., Morton, J., Podvin, S., Protsyuk, I., Purdy, T., Satterfield, K., Searles, S., Shah, S., Shires, S., Steffen, D., White, M., Todoric, J., Tuttle, R., Wojnicz, A., Sapp, V., Vargas, F., Yang, J., Zhang, C. & Dorrestein, P. C. Mass spectrometry-based visualization of molecules associated with human habitats. *Anal. Chem.* **88**, 10775–10784 (2016).

33. McCall, L.-I., Anderson, V. M., Fogle, R. S., Haffner, J. J., Hossain, E., Liu, R., Ly, A. H., Ma, H., Nadeem, M. & Yao, S. Analysis of university workplace building surfaces reveals usage-specific chemical signatures. *Build. Environ.* **162**, 106289 (2019).
34. Garg, N., Wang, M., Hyde, E., da Silva, R. R., Melnik, A. V., Protsyuk, I., Bouslimani, A., Lim, Y. W., Wong, R., Humphrey, G., Ackermann, G., Spivey, T., Brouha, S. S., Bandeira, N., Lin, G. Y., Rohwer, F., Conrad, D. J., Alexandrov, T., Knight, R. & Dorrestein, P. C. Three-Dimensional Microbiome and Metabolome Cartography of a Diseased Human Lung. *Cell Host Microbe* **22**, 705-716.e4 (2017).
35. McCall, L. I., Tripathi, A., Vargas, F., Knight, R., Dorrestein, P. C. & Siqueira-Neto, J. L. Experimental Chagas disease-induced perturbations of the fecal microbiome and metabolome. *PLoS Negl. Trop. Dis.* **12**, 1–15 (2018).
36. Dean, D. A., Haffner, J. J., Katemauswa, M. & McCall, L.-I. Chemical Cartography Approaches to Study Trypanosomatid Infection. *J. Vis. Exp.* (2022).
37. Nguyen, T., Sedghi, L., Ganther, S., Malone, E., Kamarajan, P. & Kapila, Y. Host-microbe interactions: Profiles in the transcriptome, the proteome, and the metabolome. *Physiol. Behav.* **176**, 139–148 (2017).
38. Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J. & Viant, M. R. Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9**, 44–66 (2013).
39. Gauglitz, J. M., Aceves, C. M., Aksenov, A. A., Aleti, G., Almaliti, J., Bouslimani, A., Brown, E. A., Campeau, A., Caraballo-Rodríguez, A. M., Chaar, R., da Silva, R. R., Demko, A. M., Di Ottavio, F., Elijah, E., Ernst, M., Ferguson, L. P., Holmes, X., Jarmusch, A. K., Jiang, L., Kang, K. Bin, Koester, I., Kwan, B., Li, J., Li, Y., Melnik, A. V., Molina-Santiago, C., Ni, B., Oom, A. L., Panitchpakdi, M. W., Petras, D., Quinn, R., Sikora, N., Spengler, K., Teke, B., Tripathi, A., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vrbanac, A., Vu, A. Q., Wang, S. C., Weldon, K., Wilson, K., Wozniak, J. M., Yoon, M., Bandeira, N. & Dorrestein, P. C. Untargeted mass spectrometry-based metabolomics approach unveils molecular changes in raw and processed foods and beverages. *Food Chem.* **302**, 125290 (2020).
40. Taylor, B. C., Lejzerowicz, F., Poirel, M., Shaffer, J. P., Jiang, L., Aksenov, A., Litwin, N., Humphrey, G., Martino, C., Miller-Montgomery, S., Dorrestein, P. C., Veiga, P., Song, S. J., McDonald, D., Derrien, M. & Knight, R. Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome. *mSystems* **5**, 1–16 (2020).
41. Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W. M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J. & Viant, M. R. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative

(MSI). *Metabolomics* **3**, 211–221 (2007).
42. Karu, N., Deng, L., Slae, M., Guo, A. C., Sajed, T., Huynh, H., Wine, E. & Wishart, D. S. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Anal. Chim. Acta* **1030**, 1–24 (2018).
43. McCall, L. I., Callewaert, C., Zhu, Q., Song, S. J., Bouslimani, A., Minich, J. J., Ernst, M., Ruiz-Calderon, J. F., Cavallin, H., Pereira, H. S., Novoselac, A., Hernandez, J., Rios, R., Branch, O. L. H., Blaser, M. J., Paulino, L. C., Dorrestein, P. C., Knight, R. & Dominguez-Bello, M. G. Home chemical and microbial transitions across urbanization. *Nat. Microbiol.* (2020) doi:10.1038/s41564-019-0593-4.
44. Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C. & Scalbert, A. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2017).
45. Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderon, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Boya P, C. A., TorresMendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C. ., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. Ø., Pogliano, K., Linington, R. G., Gutierrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with GNPS. *Nat. Biotechnol.* **34**, 828–837 (2016).
46. Aron, A. T., Gentry, E. C., McPhail, K. L., Nothias, L. F., Nothias-Esposito, M., Bousilmani, A., Petras, D., Gauglitz, J. M., Sikora, N., Vargas, F., van der Hooft, J. J. J., Ernst, M., Bin Kang, K., Aceves, C. M., Mauricio, C.-R. A., Koester, I., Weldon, K. C., Bertrand, S., Roullier, C., Sun, K., Tehan, R. M., Boya, C. A., Martin H., C., Gutierrez, M., Ulloa, A. M., Mora, J. A. T., Mojica-Flores, R., Lakey-Beitia, J., Vasquez-Chaves, V., Zhang, Y., Calderon, A. I., Tayler, N., Keyzers, R. A., Tugizimana, F., Ndlovu, N., Aksenov, A. A., Jarmusch, A., Schmid, R., Truman, A. W., Bandeira, N., Wang, M. & Dorrestein, P. C. Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS. *ChemRxiv* (2019).

47. Phelan, V. V. Feature-based molecular networking for metabolite annotation. in *Computational Methods and Data Analysis for Metabolomics* 227–243 (2020). doi:doi:10.1007/978-1-0716-0239-3_13.
48. Ayeni, F. A., Biagi, E., Rampelli, S., Fiori, J., Soverini, M., Audu, H. J., Cristino, S., Caporali, L., Schnorr, S. L., Carelli, V., Brigidi, P., Candela, M. & Turroni, S. Infant and Adult Gut Microbiome and Metabolome in Rural Bassa and Urban Settlers from Nigeria. *Cell Rep.* **23**, 3056–3067 (2018).
49. Kisuse, J., La-ongkham, O., Nakphaichit, M., Therdtatha, P., Momoda, R., Tanaka, M., Fukuda, S., Popluechai, S., Kespechara, K., Sonomoto, K., Lee, Y. K., Nitisinprasert, S. & Nakayama, J. Urban diets linked to gut microbiome and metabolome alterations in children: A comparative cross-sectional study in Thailand. *Front. Microbiol.* **9**, 1–16 (2018).
50. Turroni, S., Fiori, J., Rampelli, S., Schnorr, S. L., Consolandi, C., Barone, M., Biagi, E., Fanelli, F., Mezzullo, M., Crittenden, A. N., Henry, A. G., Brigidi, P. & Candela, M. Fecal metabolome of the Hadza hunter-gatherers: a host-microbiome integrative view. *Sci. Rep.* **6**, 1–9 (2016).
51. Velsko, I. M., Overmyer, K. A., Speller, C., Klaus, L., Collins, M. J., Loe, L., Frantz, L. A. F., Sankaranarayanan, K., Lewis, C. M., Martinez, J. B. R., Chaves, E., Coon, J. J., Larson, G. & Warinner, C. The dental calculus metabolome in modern and historic samples. *Metabolomics* **13**, 1–17 (2017).
52. Bonicelli, A., Mickleburgh, H. L., Chighine, A., Locci, E., Wescott, D. J. & Procopio, N. The 'ForensOMICS' approach for postmortem interval estimation from human bone by integrating metabolomics, lipidomics, and proteomics. *Elife* **11**, 1–22 (2022).
53. Hardy, K., Buckley, S., Collins, M. J., Estalrrich, A., Brothwell, D., Copeland, L., García-Tabernero, A., García-Vargas, S., De La Rasilla, M., Lalueza-Fox, C., Huguet, R., Bastir, M., Santamaría, D., Madella, M., Wilson, J., Cortés, Á. F. & Rosas, A. Neanderthal medics? Evidence for food, cooking, and medicinal plants entrapped in dental calculus. *Naturwissenschaften* **99**, 617–626 (2012).
54. Grice, E. A. & Segre, J. A. The Human Microbiome: Our Second Genome. *Annu. Rev. Genomics Hum. Genet.* **13**, 151–170 (2012).
55. Sankaranarayanan, K., Ozga, A. T., Warinner, C., Tito, R. Y., Obregon-Tito, A. J., Xu, J., Gaffney, P. M., Jervis, L. L., Cox, D., Stephens, L., Foster, M., Tallbull, G., Spicer, P. & Lewis, C. M. Gut Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. *Curr. Biol.* **25**, 3161–3169 (2015).
56. Velsko, I. M., Fellows Yates, J. A., Aron, F., Hagan, R. W., Frantz, L. A. F., Loe, L., Martinez, J. B. R., Chaves, E., Gosden, C., Larson, G. & Warinner, C. Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome* **7**, 1–20 (2019).
57. McGovern, P. E. & Hall, G. R. Charting a Future Course for Organic Residue Analysis in Archaeology. *J. Archaeol. Method Theory* **23**, 592–622 (2016).
58. Heron, C. & Evershed, R. P. The Analysis of Organic Residues and the Study of Pottery Use. *Archaeol. Method Theory* **5**, 247–284 (1993).
59. Crown, P. L., Gu, J., Hurst, W. J., Ward, T. J., Bravenec, A. D., Ali, S., Kebert, L., Berch, M., Redman, E., Lyons, P. D., Merewether, J., Phillips, D. A., Reed, L. S. & Woodson, K. Ritual drinks in the pre-Hispanic US Southwest and Mexican Northwest. *Proc. Natl. Acad. Sci.* **112**, 11436–11442 (2015).

60. Guasch-Jané, M. R., Andrés-Lacueva, C., Járegui, O. & Lamuela-Raventós, R. M. The origin of the ancient Egyptian drink Shedeh revealed using LC/MS/MS. *J. Archaeol. Sci.* **33**, 98–101 (2006).
61. Carmody, S. B., Kassabaum, M. C., Hunt, R. K., Prodanovich, N., Elliott, H. & Russ, J. Residue analysis of smoking pipe fragments from the Feltus archaeological site, Southeastern North America. *J. Archaeol. Sci. Reports* **17**, 640–649 (2018).
62. Brownstein, K. J., Tushingham, S., Damitio, W. J., Nguyen, T. & Gang, D. R. An Ancient Residue Metabolomics-Based Method to Distinguish Use of Closely Related Plant Species in Ancient Pipes. *Front. Mol. Biosci.* **7**, 1–14 (2020).
63. Zimmermann, M., Brownstein, K. J., Díaz, L. P., Aragón, I. A., Hutson, S., Kidder, B., Tushingham, S. & Gang, D. R. Metabolomics -based analysis of miniature flask contents identifies tobacco mixture use among the ancient Maya. *Sci. Rep.* 1–11 (2021) doi:10.1038/s41598-021-81158-y.
64. Garnier, N. & Vedeler, M. Revealing Medieval culinary practices in Norway: A first metabolomic-based approach. *J. Archaeol. Sci. Reports* **40**, 103206 (2021).
65. DeBerardinis, R. & Thompson, C. Cellular metabolism and disease: what do metabolic outliers teach us? **148**, 1132–1144 (2012).
66. Rappaport, S. M. Environment and Disease Risks. *Science (80-.).* **460**, 460–461 (2011).
67. Rattray, N. J. W., Deziel, N. C., Wallach, J. D., Khan, S. A., Vasiliou, V., Ioannidis, J. P. A. & Johnson, C. H. Beyond genomics : understanding exposotypes through metabolomics. 1–14 (2018) doi:10.1186/s40246-018-0134-x.
68. Lewis, C. M., McCall, L. I., Sharp, R. R. & Spicer, P. G. Ethical priority of the most actionable system of biomolecules: the metabolome. *Am. J. Phys. Anthropol.* **171**, 177–181 (2019).
69. Nordström, A., O'Maille, G., Qin, C. & Siuzdak, G. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: Quantitative analysis of endogenous and exogenous metabolites in human serum. *Anal. Chem.* **78**, 3289–3295 (2006).
70. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **15**, 473–484 (2016).
71. Lamichhane, S., Sen, P., Dickens, A. M., Orešič, M. & Bertram, H. C. Gut metabolome meets microbiome: A methodological perspective to understand the relationship between host and microbe. *Methods* **149**, 3–12 (2018).
72. Vernocchi, P., Del Chierico, F. & Putignani, L. Gut microbiota profiling: Metabolomics based approach to unravel compounds affecting human health. *Front. Microbiol.* **7**, (2016).
73. Treiman, D. J. Industrialization and Social Stratification. *Sociol. Inq.* **40**, 207–234 (1970).
74. Gollin, D., Jedwab, R. & Vollrath, D. Urbanization with and without industrialization. *J. Econ. Growth* **21**, 35–70 (2016).
75. Ritchie, H. & Roser, M. Urbanization. *Publ. online OurWorldInData.org* (2020).
76. Nguyen, T. D., Lesani, M., Forrest, I., Lan, Y., Dean, D. A., Gibaut, Q. M. R., Guo, Y., Hossain, E., Olvera, M., Panlilio, H., Parab, A. R., Wu, C., Bernatchez, J. A., Cichewicz, R. H. & McCall, L. I. Local phenomena shape backyard soil metabolite composition. *Metabolites* **10**, 1–16 (2020).
77. Höppe, P. & Martinac, I. Indoor climate and air quality. Review of current and future topics in the field of ISB study group 10. *Int. J. Biometeorol.* **42**, 1–7 (1998).
78. Katemauswa, M., Hossain, E., Liu, Z., Lesani, M., Parab, A. R., Dean, D. A. & McCall,

L.-I. Optimized extraction method enables quantitative analysis of surface metabolite recovery for exposomics and behavioral studies. 5–6 (2021).
79. Ruiz-Calderon, J. F., Cavallin, H., Song, S. J., Novoselac, A., Pericchi, L. R., Hernandez, J. N., Rios, R., Branch, O. H., Pereira, H., Paulino, L. C., Blaser, M. J., Knight, R. & Dominguez-Bello, M. G. Microbiology: Walls talk: Microbial biogeography of homes spanning urbanization. *Sci. Adv.* **2**, (2016).
80. Popkin, B. M. Urbanization, lifestyle changes and the nutrition transition. *World Dev.* **27**, 1905–1916 (1999).
81. Satterthwaite, D., McGranahan, G. & Tacoli, C. Urbanization and its implications for food and farming. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2809–2820 (2010).
82. Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T. & Lewis, C. M. Subsistence strategies in traditional societies distinguish gut microbiomes: Supplementary Information. *Nat. Commun.* **6**, 1–9 (2015).
83. Winglee, K., Howard, A. G., Sha, W., Gharaibeh, R. Z., Liu, J., Jin, D., Fodor, A. A. & Gordon-Larsen, P. Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. *Microbiome* **5**, 1–13 (2017).
84. Stamper, C. E., Hoisington, A. J., Gomez, O. M., Halweg-Edwards, A. L., Smith, D. G., Bates, K. L., Kinney, K. A., Postolache, T. T., Brenner, L. A., Rook, G. A. W. & Lowry, C. A. Chapter 14: The Microbiome of the Built Environment and Human Behavior: Implications for Emotional Health and Well-Being in Postmodern Western Societies. in *International Review of Neurobiology* 289–323 (Elsevier Inc., 2016). doi:10.1016/bs.irn.2016.07.006.
85. Carlsten, C. & Rider, C. F. Traffic-related air pollution and allergic disease: An update in the context of global urbanization. *Curr. Opin. Allergy Clin. Immunol.* **17**, 86–89 (2017).
86. Nicolaou, N., Siddique, N. & Custovic, A. Allergic disease in urban and rural populations: Increasing prevalence with increasing urbanization. *Allergy Eur. J. Allergy Clin. Immunol.* **60**, 1357–1360 (2005).
87. Ponte, E. V., Cruz, A. A., Athanazio, R., Carvalho-Pinto, R., Fernandes, F. L. A., Barreto, M. L. & Stelmach, R. Urbanization is associated with increased asthma morbidity and mortality in Brazil. *Clin. Respir. J.* **12**, 410–417 (2018).
88. Abu Freha, N., Schwartz, D., Elkrinawi, J., Ben Yakov, G., Abu Tailakh, M., Munteanu, D., Abu Ganim, A. & Fich, A. Inflammatory bowel disease among Bedouin Arabs in southern Israel: Urbanization and increasing prevalence rates. *Eur. J. Gastroenterol. Hepatol.* **27**, 230–234 (2015).
89. Barreiro-de Acosta, M., Alvarez Castro, A., Souto, R., Iglesias, M., Lorenzo, A. & Dominguez-Muñoz, J. E. Emigration to western industrialized countries: A risk factor for developing inflammatory bowel disease. *J. Crohn's Colitis* **5**, 566–569 (2011).
90. Benchimol, E. I., Kaplan, G. G., Otley, A. R., Nguyen, G. C., Underwood, F. E., Guttmann, A., Jones, J. L., Potter, B. K., Catley, C. A., Nugent, Z. J., Cui, Y., Tanyingoh, D., Mojaverian, N., Bitton, A., Carroll, M. W., Debruyn, J., Dummer, T. J. B., El-Matary, W., Griffiths, A. M., Jacobson, K., Kuenzig, M. E., Leddin, D., Lix, L. M., Mack, D. R., Murthy, S. K., Sánchez, J. N. P., Singh, H., Targownik, L. E., Vutcovici, M. & Bernstein,

C. N. Rural and Urban Residence during Early Life is Associated with a Lower Risk of Inflammatory Bowel Disease: A Population-Based Inception and Birth Cohort Study. *Am. J. Gastroenterol.* **112**, 1412–1422 (2017).
91. Bach, J. F. The hygiene hypothesis in autoimmunity: The role of pathogens and commensals. *Nat. Rev. Immunol.* **18**, 105–120 (2018).
92. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).
93. Zierer, J., Jackson, M. A., Kastenmüller, G., Mangino, M., Long, T., Telenti, A., Mohney, R. P., Small, K. S., Bell, J. T., Steves, C. J., Valdes, A. M., Spector, T. D. & Menni, C. The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* **50**, 790–795 (2018).
94. Claesson, M. J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., De Weerd, H., Flannery, E., Marchesi, J. R., Falush, D., Dinan, T., Fitzgerald, G., Stanton, C., Van Sinderen, D., O'Connor, M., Harnedy, N., O'Connor, K., Henry, C., O'Mahony, D., Fitzgerald, A. P., Shanahan, F., Twomey, C., Hill, C., Ross, R. P. & O'Toole, P. W. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4586–4591 (2011).
95. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R. & Gordon, J. I. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
96. Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T. & Lewis, C. M. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 1–9 (2015).
97. Want, E. J., Masson, P., Michopoulos, F., Wilson, I. D., Theodoridis, G., Plumb, R. S., Shockcor, J., Loftus, N., Holmes, E. & Nicholson, J. K. Global metabolic profiling of animal and human tissues via UPLC-MS. *Nat. Protoc.* **8**, 17–32 (2013).
98. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L. & Mallick, P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
99. Nothias, L. F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., Protsyuk, I., Ernst, M. & Tsugawa, H. Feature-based Molecular Networking in the GNPS Analysis Environment. *biorXiv* (2019).
100. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, (2010).
101. Bunn, A. & Korpela, M. R: A Language for Data Analysis and Graphics. *Ind. Commer. Train.* **0**, 1–16 (2018).

102. Kluyver, T., Ragan-kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S. & Willing, C. Jupyter Notebooks—a publishing format for reproducible computational workflows. *Position. Power Acad. Publ. Play. Agents Agendas* 87–90 (2016) doi:10.3233/978-1-61499-649-1-87.
103. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schikowski, B. & Ideker, T. Cytoscape: A Software Environment for Integrated Models. *Genome Res.* **13**, 2498–2504 (2003).
104. Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., Greiner, R. & Wishart, D. S. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 1–20 (2016).
105. Ernst, M., Kang, K. Bin, Caraballo-Rodríguez, A. M., Nothias, L.-F., Wandy, J., Chen, C., Wang, M., Rogers, S., Medema, M. H., Dorrestein, P. C. & van der Hooft, J. J. J. MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* **9**, 144 (2019).
106. Wang, M., Jarmusch, A. K., Vargas, F., Aksenov, A. A., Gauglitz, J. M., Weldon, K., Petras, D., da Silva, R., Quinn, R., Melnik, A. V., van der Hooft, J. J. J., Caraballo-Rodríguez, A. M., Nothias, L. F., Aceves, C. M., Panitchpakdi, M., Brown, E., Di Ottavio, F., Sikora, N., Elijah, E. O., Labarta-Bajo, L., Gentry, E. C., Shalapour, S., Kyle, K. E., Puckett, S. P., Watrous, J. D., Carpenter, C. S., Bouslimani, A., Ernst, M., Swafford, A. D., Zúñiga, E. I., Balunas, M. J., Klassen, J. L., Loomba, R., Knight, R., Bandeira, N. & Dorrestein, P. C. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 19–22 (2020).
107. Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., Sumner, S. & Subramaniam, S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
108. Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S. A., Griffin, J. L. & Steinbeck, C. MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, 781–786 (2013).
109. Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L. O., Draper, J., Rappaport, S. M., Van Der Hooft, J. J. J. & Wishart, D. S. The food metabolome: A window over dietary exposure. *Am. J. Clin. Nutr.* **99**, 1286–1308 (2014).
110. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
111. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
112. Chun, J., Lee, J. H., Jung, Y., Kim, M., Kim, S., Kim, B. K. & Lim, Y. W. EzTaxon: A

web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int. J. Syst. Evol. Microbiol.* **57**, 2259–2261 (2007).
113. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G. L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
114. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. Bin, Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y. X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. & Caporaso, J. G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
115. Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Van Goethem, M. W., Northen, T. R., Vazquez-Baeza, Y., Wang, M., Bokulich, N. A., Watters, A., Song, S. J., Bonneau, R., Dorrestein, P. C. & Knight, R. Learning representations of microbe–metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).
116. Weiner, J. Package 'pca3d': Three Dimensional PCA Plots. (2020).
117. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: A tool for visualizing high-throughput microbial community data. *Gigascience* **2**, 2–5 (2013).
118. Anderson, M. J. Permutational Multivariate Analysis of Variance (PERMANOVA) . *Wiley StatsRef Stat. Ref. Online* 1–15 (2017) doi:10.1002/9781118445112.stat07841.
119. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer US, 2016).
120. South, A. rworldmap: A new R package for mapping global data. *R J.* **3**, 35–43 (2011).
121. Ben-Shachar, M., Makowski, D. & Lüdecke, D. Compute and interpret indices of effect size. *CRAN* (2020).
122. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
123. van Rossum, G. & Drake, F. L. Python 3 Reference Manual. at https://www.python.org/ (2009).
124. McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python high Perform. Sci. Comput.* **14**, 1–9 (2011).

125. Nothman, J. UpSetPlot documentation. at https://upsetplot.readthedocs.io/en/stable/ (2018).
126. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
127. Liaw, A. & Wiener, M. Classification and Regression by RandomForest. *R News* **3**, 18–22 (2002).
128. Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., Dorrestein, P. C., Rousu, J. & Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
129. Dührkop, K., Nothias, L.-F., Fleischauer, M., Ludwig, M., Hoffmann, M. A., Rousu, J., Dorrestein, P. C. & Böcker, S. Classes for the masses: Systematic classification of unknowns using fragmentation spectra. *bioRxiv* 2020.04.17.046672 (2020) doi:10.1101/2020.04.17.046672.
130. Schnorr, S. L., Mandela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., De Bellis, G., Luiselli, D., Brigidi, P., Mabulla, A., Marlowe, F., Henry, A. G. & Crittenden, A. N. Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 1–12 (2014).
131. Gomez, A., Petrzelkova, K. J., Burns, M. B., Yeoman, C. J., Amato, K. R., Vlckova, K., Modry, D., Todd, A., Jost Robinson, C. A., Remis, M. J., Torralba, M. G., Morton, E., Umaña, J. D., Carbonero, F., Gaskins, H. R., Nelson, K. E., Wilson, B. A., Stumpf, R. M., White, B. A., Leigh, S. R. & Blekhman, R. Gut Microbiome of Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns. *Cell Rep.* **14**, 2142–2153 (2016).
132. Mancabelli, L., Milani, C., Lugli, G. A., Turroni, F., Ferrario, C., van Sinderen, D. & Ventura, M. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environ. Microbiol.* **19**, 1379–1390 (2017).
133. Wu, G. D., Compher, C., Chen, E. Z., Smith, S. A., Shah, R. D., Bittinger, K., Chehoud, C., Albenberg, L. G., Nessel, L., Gilroy, E., Star, J., Weljie, A. M., Flint, H. J., Metz, D. C., Bennett, M. J., Li, H., Bushman, F. D. & Lewis, J. D. Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. *Gut* **65**, 63–72 (2016).
134. Meghwal, M. & Goswami, T. K. Piper nigrum and piperine: An update. *Phyther. Res.* **27**, 1121–1130 (2013).
135. Jandke, J. & Spiteller, G. Dipeptide analysis in human urine. *J. Chromatogr. B Biomed. Sci. Appl.* **382**, 39–45 (1986).
136. Mattarucchi, E., Baraldi, E. & Guillou, C. Metabolomics applied to urine samples in childhood asthma; differentiation between asthma phenotypes and identification of relevant metabolites. *Biomed. Chromatogr.* **26**, 89–94 (2011).
137. Jones-Lepp, T. L. Chemical markers of human waste contamination: Analysis of urobilin and pharmaceuticals in source waters. *J. Environ. Monit.* **8**, 472–478 (2006).
138. Cao, H., Huang, H., Xu, W., Chen, D., Yu, J., Li, J. & Li, L. Fecal metabolome profiling of liver cirrhosis and hepatocellular carcinoma patients by ultra performance liquid chromatography-mass spectrometry. *Anal. Chim. Acta* **691**, 68–75 (2011).
139. Jarmusch, A. K., Wang, M., Aceves, C. M., Advani, R. S., Aguire, S., Aksenov, A. A., Aleti, G., Aron, A. T., Bauermeister, A., Bolleddu, S., Bouslimani, A., Rodriguez, A. M. C., Chaar, R., Coras, R., Elijah, E. O., Ernst, M., Gauglitz, J. M., Gentry, E. C., Husband, M., Jarmusch, S. A., Jones, K. L., Kamenik, Z., Gouellec, A. Le, Lu, A., McCall, L.-I.,

McPhail, K. L., Meehan, M. J., Melnik, A. V., Menezes, R. C., Giraldo, Y. A. M., Nguyen, N. H., Nothias, L. F., Nothias-Esposito, M., Panitchpakdi, M., Petras, D., Quinn, R., Sikora, N., Hooft, J. J. J. van der, Vargas, F., Vrbanac, A., Weldon, K., Knight, R., Bandeira, N. & Dorrestein, P. C. Repository-scale Co- and Re-analysis of Tandem Mass Spectrometry Data. *bioRxiv* 750471 (2019) doi:10.1101/750471.

140. Hill, M. J. The effect of some factors on the faecal concentration of acid steroids, neutral steroids and urobilins. *J. Pathol.* **104**, 239–245 (1971).
141. Jain, A., Li, X. H. & Chen, W. N. An untargeted fecal and urine metabolomics analysis of the interplay between the gut microbiome, diet and human metabolism in Indian and Chinese adults. *Sci. Rep.* **9**, 1–13 (2019).
142. Requena, T., Martínez-Cuesta, M. C. & Peláez, C. Diet and microbiota linked in health and disease. *Food Funct.* **9**, 688–704 (2018).
143. Tang, Z. Z., Chen, G., Hong, Q., Huang, S., Smith, H. M., Shah, R. D., Scholz, M. & Ferguson, J. F. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front. Genet.* **10**, (2019).
144. Hoffmann, M. A., Nothias, L.-F., Ludwig, M., Fleischauer, M., Gentry, E. C., Witting, M., Dorrestein, P. C., Dührkop, K. & Böcker, S. Assigning confidence to structural annotations from mass spectra with COSMIC. *bioRxiv* 2021.03.18.435634 (2021).
145. Segata, N. Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. *Curr. Biol.* **25**, R611–R613 (2015).
146. Katajamaa, M. & Orešič, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* **6**, (2005).
147. Pluskal, T., Korf, A., Smirnov, A., Schmid, R., Fallon, T. R., Du, X. & Weng, J. K. Chapter 7: Metabolomics Data Analysis Using MZmine. in *Processing Metabolomics and Proteomics Data with Open Software* 232–254 (The Royal Society of Chemistry, 2020). doi:10.1039/9781788019880-00232.
148. Chiang, J. Y. L. Bile acid metabolism and signaling in liver disease and therapy. *Liver Res.* **1**, 3–9 (2017).
149. Ridlon, J. M., Kang, D. J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *J. Lipid Res.* **47**, 241–259 (2006).
150. Chickos, J. S., Way, B. A., Wilson, J., Shaharuzzaman, M., Laird, J. & Landt, M. Analysis of 3-hydroxydodecanedioic acid for studies of fatty acid metabolic disorders: Preparation of stable isotope standards. *J. Clin. Lab. Anal.* **16**, 115–120 (2002).
151. Ktsoyan, Z. A., Beloborodova, N. V., Sedrakyan, A. M., Osipov, G. A., Khachatryan, Z. A., Kelly, D., Manukyan, G. P., Arakelova, K. A., Hovhannisyan, A. I., Olenin, A. Y., Arakelyan, A. A., Ghazaryan, K. A. & Aminov, R. I. Profiles of microbial fatty acids in the human metabolome are disease-specific. *Front. Microbiol.* **1**, 1–9 (2011).
152. Ogunleye, A., Fakoya, A. T., Niizeki, S., Tojo, H., Sasajima, I., Kobayashi, M., Tateishi, S. & Yamaguchi, K. Fatty Acid Composition of Breast Milk from Nigerian and Japanese Women. *J. Nutr. Sci. Vitaminol. (Tokyo).* **37**, 435–442 (1991).
153. Ozalp, A., Barroso, B., Meijer, J. & van den Beld, C. Determination of methionine-enkephalin and leucine-enkephalin by LC-MS in human plasma: Study of pre-analytical stability. *Anal. Biochem.* **559**, 24–29 (2018).
154. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).

155. Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H. & Adam, T. PLS-DA for compositional data with application to metabolomics. *J. Chemom.* **29**, 21–28 (2015).
156. Lewis, C. M., Obregón-Tito, A., Tito, R. Y., Foster, M. W. & Spicer, P. G. The Human Microbiome Project: Lessons from human genomics. *Trends Microbiol.* **20**, 1–4 (2012).
157. Makki, K., Deehan, E. C., Walter, J. & Bäckhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host Microbe* **23**, 705–715 (2018).
158. Mishra, S. P., Karunakar, P., Taraphder, S. & Yadav, H. Free Fatty Acid Receptors 2 and 3 as Microbial Metabolite Sensors to Shape Host Health: Pharmacophysiological View. *Biomedicines* **8**, 1–45 (2020).
159. Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, L., Griffi, N. W., Lombard, V., Henrissat, B., Bain, J. R., Michael, J., Ilkayeva, O., Semenkovich, C. F., Funai, K., Hayashi, D. K., Lyle, J., Martini, M. C., Ursell, L. K., Clemente, J. C., Treuren, W. Van, William, A., Knight, R., Newgard, C. B., Heath, A. C., Gordon, J. I., Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., Griffin, N. W., Lombard, V., Henrissat, B., Bain, J. R., Muehlbauer, M. J., Ilkayeva, O., Semenkovich, C. F., Funai, K., Hayashi, D. K., Lyle, B. J., Martini, M. C., Ursell, L. K., Clemente, J. C., Van Treuren, W., Walters, W. A., Knight, R., Newgard, C. B., Heath, A. C. & Gordon, J. I. Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice Gut Microbiota from Twins Metabolism in Mice. *Science (80-.).* **341**, 1241214-1–9 (2013).
160. Cao, Y., Aquino-martinez, R., Hutchison, E., Allayee, H., Lusis, A. J. & Rey, F. E. Role of gut microbe-derived metabolites in cardiometabolic diseases: Systems based approach. *Mol. Metab.* 101557 (2022) doi:10.1016/j.molmet.2022.101557.
161. Anderson, J. W., Smith, B. M. & Gustafson, N. J. Health benefits and practical of high-fiber diets. *Am. J. Clin. Nutr.* **59**, 1242S-1247S (1994).
162. Câmara, C. R., Urrea, C. A. & Schlegel, V. Pinto Beans (Phaseolus vulgaris L.) as a Functional Food: Implications on Human Health. *Agriculture* **3**, 90–111 (2013).
163. Gong, L., Wang, T., Sun, C., Wang, J. & Sun, B. Whole barley prevents obesity and dyslipidemia without the involvement of the gut microbiota in germ free C57BL/6J obese mice. *Food Funct.* **10**, 7498–7508 (2019).
164. Afshin, A., Micha, R., Khatibzadeh, S. & Mozaffarian, D. Consumption of nuts and legumes and risk of incident ischemic heart disease, stroke, and diabetes: a systematic review and meta-analysis. *Am. J. Clin. Nutr.* **100**, 278–288 (2014).
165. Andersson, A., Tengblad, S., Karlstro, B., Kamal-eldin, A., Landberg, R., Basu, S., Per, Å. & Vessby, B. Whole-Grain Foods Do Not Affect Insulin Sensitivity or Markers of Lipid Peroxidation and Inflammation in Healthy, Moderately Overweight Subjects. *J. Nutr.* **137**, 1401–1407 (2007).
166. Brownlee, I. A., Moore, C., Chatfield, M., Richardson, D. P., Ashby, P., Kuznesof, S. A., Jebb, S. A. & Seal, C. J. Markers of cardiovascular risk are not changed by increased whole-grain intake: the WHOLEheart study, a randomised, controlled dietary intervention. *Br. J. Nutr.* **104**, 125–134 (2010).
167. Martıínez, I., Lattimer, J. M., Hubach, K. L., Case, J. A., Yang, J., Weber, C. G., Louk, J. A., Rose, D. J., Kyureghian, G., Peterson, D. A., Haub, M. D. & Walter, J. Gut microbiome composition is linked to whole grain-induced immunological improvements. *ISME J.* **7**, 269–280 (2013).
168. Fujihara, K., Nogawa, S., Saito, K., Horikawa, C., Takeda, Y., Cho, K., Ishiguro, H., Kodama, S., Nakagawa, Y., Matsuzaka, T., Shimano, H. & Sone, H. Carrot Consumption

Frequency Associated with Reduced BMI and Obesity through the SNP Intermediary rs4445711. *Nutrients* **13**, 1–11 (2021).
169. Bendiks, Z. A., Knudsen, K. E. B., Keenan, M. J. & Marco, M. L. Conserved and variable responses of the gut microbiome to resistant starch type 2. *Nutr. Res.* **77**, 12–28 (2020).
170. Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., Avnit-Sagi, T., Kosower, N., Malka, G., Rein, M., Suez, J., Goldberg, B. Z., Weinberger, A., Levy, A. A., Elinav, E. & Segal, E. Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab.* **25**, 1243-1253.e5 (2017).
171. Gowd, V., Xie, L., Zheng, X. & Chen, W. Dietary fibers as emerging nutritional factors against diabetes: focus on the involvement of gut microbiota. *Crit. Rev. Biotechnol.* **39**, 524–540 (2019).
172. Lovegrove, A., Edwards, C. H., Noni, I. De, Patel, H., El, S. N., Grassby, T., Zielke, C., Ulmius, M., Nilsson, L., Butterworth, P. J., Ellis, P. R. & Shewry, P. R. Role of polysaccharides in food, digestion, and health. *Crit. Rev. Food Sci. Nutr.* **57**, 237–253 (2017).
173. Gill, S. K., Rossi, M., Bajka, B. & Whelan, K. Dietary fibre in gastrointestinal health and disease. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 101–116 (2021).
174. Esmaillzadeh, A., Mirmiran, P. & Azizi, F. Whole-grain consumption and the metabolic syndrome: a favorable association in Tehranian adults. *Eur. J. Clin. Nutr.* **59**, 353–362 (2005).
175. Chiva-blanch, G. & Badimon, L. Effects of Polyphenol Intake on Metabolic Syndrome: Current Evidences from Human Trials. *Oxid. Med. Cell. Longev.* **2017**, 1–17 (2017).
176. Utrilla, M. P., Peinado, M. J., Ruiz, R., Rodriguez-Nogales, A., Algieri, F., Rodriguez-cabezas, M. E., Clemente, A., Galvez, J. & Rubio, L. A. Pea (Pisum sativum L.) seed albumin extracts show anti-inflammatory effect in the DSS model of mouse. *Mol. Nutr. Food Res.* **59**, 807–819 (2015).
177. Aranda-Olmedo, I., Ruiz, R., Peinado, M. J. & Rubio, L. A. A pea (Pisum sativum L.) seed albumin extract prevents colonic DSS induced dysbiosis in mice. *J. Funct. Foods* **35**, 279–294 (2017).
178. Benezra, A., DeStefano, J. & Gordon, J. I. Anthropology of microbes. *Proc. Natl. Acad. Sci.* **109**, 6378–6381 (2012).
179. Herd, P., Carr, D. & Roan, C. Cohort profile: Wisconsin longitudinal study (WLS). *Int. J. Epidemiol.* **43**, 34–41 (2014).
180. Murga-Garrido, S. M., Hong, Q., Cross, T. L., Hutchison, E. R., Han, J., Thomas, S. P., Vivas, E. I., Denu, J., Ceschin, D. G., Tang, Z. & Rey, F. E. Gut microbiome variation modulates the effects of dietary fiber on host metabolism. *Microbiome* **9**, 1–26 (2021).
181. Li, B. W., Andrews, K. W. & Pehrsson, P. R. Individual Sugars, Soluble, and Insoluble Dietary Fiber Contents of 70 High Consumption Foods. *J. Food Compos. Anal.* **15**, 715–723 (2002).
182. Lucier, G., Lin, B. H., Allshouse, J. & Kantor, L. S. Factors affecting dry bean consumption in the United States. *Small* **19**, 2–5 (2000).
183. Repo-Carrasco-Valencia, R. A.-M. & Serna, L. A. Quinoa (Chenopodium quinoa, Willd.) as a source of dietary fiber and other functional components. *Food Sci. Technol.* **31**, 225–230 (2011).
184. Ichimura, M., Kawase, M., Masuzumi, M., Sakaki, M., Nagata, Y., Tanaka, K., Suruga, K., Tamaru, S., Kato, S., Tsuneyama, K. & Omagari, K. High-fat and high-cholesterol diet

rapidly induces non-alcoholic steatohepatitis with advanced fibrosis in S prague–D awley rats. *Hepatol. Res.* **45**, 458–469 (2015).
185. Kratz, M., Baars, T. & Guyenet, S. The relationship between high-fat dairy consumption and obesity, cardiovascular, and metabolic disease. *Eur. J. Nutr.* **52**, 1–24 (2013).
186. McCall, L.-I., Morton, J. T., Bernatchez, J. A., Siqueira-neto, J. L. De, Knight, R., Dorrestein, P. C. & McKerrow, J. H. Mass Spectrometry-Based Chemical Cartography of a Cardiac Parasitic Infection. *Anal. Chem.* **89**, 10414–10421 (2017).
187. Hossain, E., Khanam, S., Dean, D. A., Wu, C., Lostracco-Johnson, S., Thomas, D., Kane, S. S., Parab, A. R., Flores, K., Katemauswa, M., Gosmanov, C., Hayes, S. E., Zhang, Y., Li, D., Woelfel-Monsivais, C., Sankaranarayanan, K. & McCall, L.-I. Mapping of host-parasite-microbiome interactions reveals metabolic determinants of tropism and tolerance in Chagas disease. *Sci. Adv.* **6**, 1–15 (2020).
188. Hoffman, K., Liu, Z., Hossain, E., Bottazzi, M. E., Hotez, P. J., Jones, K. M. & McCall, L. Alterations to the Cardiac Metabolome Induced by Chronic T. cruzi Infection Relate to the Degree of Cardiac Pathology. *ACS Infect. Dis.* **7**, 1638–1649 (2021).
189. Dean, D. A., Klechka, L., Hossain, E., Parab, A. R., Eaton, K., Hinsdale, M. & McCall, L.-I. Spatial Metabolomics Reveals Localized Impact of Influenza Virus Infection on the Lung Tissue Metabolome. *mSystems* **7**, (2022).
190. Nothias, L. F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., Protsyuk, I., Ernst, M., Tsugawa, H., Fleischauer, M., Aicheler, F., Aksenov, A. A., Alka, O., Allard, P. M., Barsch, A., Cachet, X., Caraballo-Rodriguez, A. M., Da Silva, R. R., Dang, T., Garg, N., Gauglitz, J. M., Gurevich, A., Isaac, G., Jarmusch, A. K., Kameník, Z., Kang, K. Bin, Kessler, N., Koester, I., Korf, A., Le Gouellec, A., Ludwig, M., Martin H, C., McCall, L. I., McSayles, J., Meyer, S. W., Mohimani, H., Morsy, M., Moyne, O., Neumann, S., Neuweger, H., Nguyen, N. H., Nothias-Esposito, M., Paolini, J., Phelan, V. V., Pluskal, T., Quinn, R. A., Rogers, S., Shrestha, B., Tripathi, A., van der Hooft, J. J. J., Vargas, F., Weldon, K. C., Witting, M., Yang, H., Zhang, Z., Zubeil, F., Kohlbacher, O., Böcker, S., Alexandrov, T., Bandeira, N., Wang, M. & Dorrestein, P. C. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
191. Chen, L., Lu, W., Wang, L., Xing, X., Chen, Z., Teng, X., Zeng, X., Muscarella, A. D., Shen, Y., Cowan, A., Mcreynolds, M. R., Kennedy, B. J., Lato, A. M., Campagna, S. R., Singh, M. & Rabinowitz, J. D. Metabolite discovery through global annotation of untargeted metabolomics data. *Nat. Methods* **18**, 1377–1385 (2021).
192. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 1–4 (2021).
193. Harris, C. R., Millman, K. J., Walt, S. J. Van Der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. Van, Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
194. Rechner, A. R., Smith, M. A., Kuhnle, G., Gibson, G. R., Debnam, E. S., Srai, S. K. S., Moore, K. P. & Rice-Evans, C. A. Colonic metabolism of dietary polyphenols: influence of structure on microbial fermentation products. *Free Radic. Biol. Med.* **36**, 212–225 (2004).

195. Cohen, L. J., Kang, H., Chu, J., Huang, Y., Gordon, E. A., Reddy, B. V. B., Ternei, M. A., Craig, J. W. & Brady, S. F. Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist. *Proc. Natl. Acad. Sci.* **112**, E4825–E4834 (2015).
196. Agus, A., Planchais, J. & Sokol, H. Gut microbiota regulation of tryptophan metabolism in health and disease. *Cell Host Microbe* **23**, 716–724 (2018).
197. Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C. & Siuzdak, G. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci.* **106**, 3698–3703 (2009).
198. González-Sarrías, Espín, J. C. & Tomás-Barberán, F. A. Non-extractable polyphenols produce gut microbiota metabolites that persist in circulation and show anti-inflammatory and free radical-scavenging effects. *Trends Food Sci. Technol.* **69**, 281–288 (2017).
199. Barker, A., Dombrosky, J., Chaput, D., Venables, B., Wolverton, S. & Stevens, S. M. Validation of a Non - Targeted LC - MS Approach for Identifying Ancient Proteins : Method Development on Bone to Improve Artifact Residue Analysis. *Ethnobiol. Lett.* **6**, 162–174 (2015).
200. Parnell, J. J. & Terry, R. E. Soil Chemical Analysis of Ancient Activities in Cerén, El Salvador: A Case Study of a Rapidly Abandoned Site. *Lat. Am. Antiq.* **13**, 331–342 (2002).
201. Porru, E., Giorgi, E., Turroni, S., Helg, R., Silani, M., Candela, M., Fiori, J. & Roda, A. Bile acids and oxo-metabolites as markers of human faecal input in the ancient Pompeii ruins. *Sci. Rep.* **11**, 1–7 (2021).
202. Eerkens, J. W., Tushingham, S., Brownstein, K. J., Garibay, R., Perez, K., Murga, E., Kaijankoski, P., Rosenthal, S. & Gang, D. R. Dental calculus as a source of ancient alkaloids : Detection of nicotine by LC- MS in calculus samples from the Americas. *J. Archaeol. Sci. Reports* **18**, 509–515 (2018).
203. Barker, A., Dombrosky, J., Venables, B. & Wolverton, S. Taphonomy and negative results: An integrated approach to ceramic-bound protein residue analysis. *J. Archaeol. Sci.* **94**, 32–43 (2018).
204. Crown, P. L. & Hurst, W. J. Evidence of cacao use in the Prehispanic American Southwest. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2110–2113 (2009).
205. Washburn, D. K., Washburn, W. N., Shipkova, P. A. & Pelleymounter, M. A. Chemical analysis of cacao residues in archaeological ceramics from North America: Considerations of contamination, sample size and systematic controls. *J. Archaeol. Sci.* **50**, 191–207 (2014).
206. Roffet-Salque, M., Dunne, J., Altoft, D. T., Casanova, E., Cramp, L. J. E., Smyth, J., Whelton, H. L. & Evershed, R. P. From the inside out: Upscaling organic residue analyses of archaeological ceramics. *J. Archaeol. Sci. Reports* **16**, 627–640 (2017).
207. King, A., Powis, T. G., Cheong, K. F., Deere, B., Pickering, R. B., Singleton, E. & Gaikwad, N. W. Absorbed Residue Evidence for Prehistoric Datura Use in the American Southeast and Western Mexico. *Adv. Archaeol. Pract.* **6**, 312–327 (2018).
208. Tushingham, S., Snyder, C. M., Brownstein, K. J., Damitio, W. J. & Gang, D. R. Biomolecular archaeology reveals ancient origins of indigenous tobacco smoking in North American Plateau. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 11742–11747 (2018).
209. Marley, M. B., Brownstein, K. J., Lepper, B. T., Tushingham, S., Leone, K. L., Rafferty, S. M., Gang, D. R., Pansing, L. L., Pickard, W. H. & Six, J. Analyses of organic residue

from a conical pipe from the Niles-Wolford Mound (33Pi3), Pickaway County, Ohio. *J. Archaeol. Sci. Reports* **19**, 658–668 (2018).
210. Alder, L., Greulich, K., Kempe, G. & Vieth, B. Residue Analysis of 500 High Priority Pesticides: Better by GC-MS or LC-MS/MS? *Mass Spectrom. Rev.* **25**, 838–865 (2006).
211. Perez, E. R., Knapp, J. A., Horn, C. K., Stillman, S. L., Evans, J. E. & Arfsten, D. P. Comparison of LC-MS-MS and GC-MS analysis of benzodiazepine compounds included in the drug demand reduction urinalysis program. *J. Anal. Toxicol.* **40**, 201–207 (2016).
212. Drieu, L., Peche-Quilichini, K., Lachenal, T. & Regert, M. Domestic activities and pottery use in the Iron Age Corsican settlement of Cuciurpula revealed by organic residue analysis. *J. Archaeol. Sci. Reports* **19**, 213–223 (2018).
213. García-Granero, J. J., Suryanarayan, A., Cubas, M., Craig, O. E., Cárdenas, M., Ajithprasad, P. & Madella, M. Integrating Lipid and Starch Grain Analyses From Pottery Vessels to Explore Prehistoric Foodways in Northern Gujarat, India. *Front. Ecol. Evol.* **10**, (2022).
214. Koh, A. J., Betancourt, P. P., Pareja, M. N., Brogan, T. M. & Apostolakou, V. Organic residue analysis of pottery from the dye workshop at Alatsomouri-Pefka, Crete. *J. Archaeol. Sci. Reports* **7**, 536–538 (2016).
215. Miller, M. J., Whelton, H. L., Swift, J. A., Maline, S., Hammann, S., Cramp, L. J. E., McCleary, A., Taylor, G., Vacca, K., Becks, F., Evershed, R. P. & Hastorf, C. A. Interpreting ancient food practices: stable isotope and molecular analyses of visible and absorbed residues from a year-long cooking experiment. *Sci. Rep.* **10**, 1–17 (2020).
216. Qanbari-Taheri, N., Karimy, A. H., Holakooei, P. & Kobarfard, F. Organic residue analysis of Iron Age ceramics from the archaeological site of Kani-zirin, western Iran. *Archaeometry* **62**, 612–625 (2020).
217. Kaplan, J., Paredes Umaña, F., Hurst, W. J., Sun, D., Stanley, B., Barba Pingarrón, L. & Obregon Cardona, M. Cacao residues in vessels from Chocolá, an early Maya polity in the southern Guatemalan piedmont, determined by semi-quantitative testing and high-performance liquid chromatography. *J. Archaeol. Sci. Reports* **13**, 526–534 (2017).
218. Technology, N. I. of S. and. NIST/EPA/NIH Mass Spectral Library. *US Secr. Commer.*
219. Warinner, C., Robles, N. & Tuross, N. Maize, beans and the floral isotopic diversity of highland Oaxaca, Mexico. *J. Archaeol. Sci.* **40**, 868–873 (2013).
220. Ashihara, H. & Suzuki, T. Distribution and biosynthesis of caffeine in plants. *Front. Biosci.* **9**, 1864–1876 (2004).
221. Chapa-Oliver, A. M. & Mejía-Teniente, L. Capsaicin: From Plants to a Cancer-Suppressing Agent. *Molecules* **21**, 1–14 (2016).
222. Tewksbury, J. J. & Nabhan, G. P. Directed deterrence by capsaicin in chillies. *Nature* **412**, 403–404 (2001).
223. Calderón-Montaño, J. M., Burgos-Morón, E., Pérez-Guerrero, C. & López-Lázaro, M. A Review on the Dietary Flavonoid Kaempferol. *Mini-Reviews Med. Chem.* **11**, 298–344 (2011).
224. Bolarinwa, I. F., Orfila, C. & Morgan, M. R. A. Determination of amygdalin in apple seeds, fresh apples and processed apple juices. *Food Chem.* **170**, 437–442 (2015).
225. Jaszczak-Wilke, E., Polkowska, Ż., Koprowski, M., Owsianik, K., Mitchell, A. E. & Bałczewski, P. Amygdalin: Toxicity, Anticancer Activity and Analytical Procedures for Its Determination in Plant Seeds. *Molecules* **26**, 1–16 (2021).
226. Rasul, A., Millimouno, F. M., Eltayb, W. A., Ali, M., Li, J. & Li, X. Pinocembrin: A

Novel Natural Compound with Versatile Pharmacological and Biological Activities. *Biomed Res. Int.* 1–10 (2013).
227. Jaganathan, S. K. & Mandal, M. Antiproliferative Effects of Honey and of Its Polyphenols: A Review Chemistry of Honey. *J. Biomed. Biotechnol.* 1–14 (2009) doi:10.1155/2009/830616.
228. Kwik-Uribe, C. & Bektash, R. M. Cocoa flavanols: measurement, bioavailability and bioactivity. *Asia Pac. J. Clin. Nutr.* **17**, 280–283 (2008).
229. Liu, Y., Gao, L., Liu, L., Yang, Q., Lu, Z., Nie, Z., Wang, Y. & Xia, T. Purification and Characterization of a Novel Galloyltransferase Involved in Catechin Galloylation in the Tea Plant (Camellia sinensis)*. *J. Biol. Chem.* **287**, 44406–44417 (2012).
230. Athilli, L. & Balasubramaniam, A. Hordenine: Pharmacological, phytochemical, pharmacokinetic and analytical review of the literature. *Ann. Phytomedicine* **11**, 121–132 (2022).
231. Gregor, R., Probst, M., Eyal, S., Aksenov, A., Sasson, G., Horovitz, I., Dorrestein, P. C., Meijler, M. M. & Mizrahi, I. Mammalian gut metabolomes mirror microbiome composition and host phylogeny. *ISME J.* **16**, 1262–1274 (2022).
232. Timm, C. M., Loomis, K., Stone, W., Mehoke, T., Brensinger, B., Pellicore, M., Staniczenko, P. P. A., Charles, C., Nayak, S. & Karig, D. K. Isolation and characterization of diverse microbial representatives from the human skin microbiome. *Microbiome* **8**, 1–12 (2020).

Supplementary Material A

Author List and Affiliations

Jacob J. Haffner[1,2], Mitchelle Katemauswa[2,3], Thérèse S. Kagone[4,5], Ekram Hossain[2,3], David Jacobson[1,2], Karina Flores[2,6], Adwaita R. Parab[2,7], Alexandra J. Obregon-Tito[1,2], Raul Y. Tito[1,2], Luis Marin Reyes[8], Luzmila Troncoso-Corzo[9], Emilio Guija-Poma[10], Nicolas Meda[4], Hélène Carabin[11,12,13,14], Tanvi P. Honap[1,2], Krithivasan Sankaranarayanan[2,7], Cecil M. Lewis Jr[1,2*], Laura-Isobel McCall[2,3,7*]

[1]Department of Anthropology, University of Oklahoma, Norman, OK, USA.

[2]Laboratories of Molecular Anthropology and Microbiome Research (LMAMR), University of Oklahoma, Norman, OK.

[3]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA.
[4]Burkina Faso Ministry of Health

[5]Centre MURAZ Research Institute, Bobo-Dioulasso, Burkina Faso.

[6]Department of Biology, University of Oklahoma, Norman, OK, USA.

[7]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA.

[8]Instituto Nacional de Salud, Lima, Peru.

[9]Facultad de Medicina, Universidad Nacional Mayor de San Marcos, Lima, Peru.

[10]Centro de Investigación de Bioquímica y Nutrición, Facultad de Medicina Humana, Universidad de San Martín de Porres, Lima, Perú.

[11]Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA.

[12]Département de Pathologie et Microbiologie, Faculté de médecine vétérinaire, Université de Montréal, Saint-Hyacinthe, QC J2S 2M2, Canada.

[13]Département de médecine sociale et préventive, École de santé publique de l'université de Montréal, Montréal, QC H3N 1X9, Canada.

[14]Centre de Recherche en Santé Publique (CReSP) de l'université de Montréal et du CIUSS du Centre Sud de Montréal, Montréal, QC H3N 1X9, Canada.

*Corresponding author(s)

Author Contributions

C.M.L., L.-I.M., and K.S. conceived and designed the study. C.M.L., A.J.O.-T., R.Y.T., L.M.-R., E.G.-P., and L.T.-C. led Peruvian sample collection and developed ethical guidelines

for community engagement. T.S.K. led fieldwork, metadata curation, and sample processing in Burkina Faso and contributed to lab work in the U.S. D.J. assisted with fieldwork in Burkina Faso, metadata curation, and conducted data analysis. L.-I.M. directed all LC-MS/MS experimentation and data analyses. J.J.H., E.H., and L.-I.M. acquired LC-MS/MS data. J.J.H. and L-I.M. performed LC-MS/MS data analysis with contributions from M.K., A.R.P., and K.F. J.J.H. wrote the manuscript with contributions from L.-I.M. and C.M.L. All authors reviewed the final manuscript.

Acknowledgements

We thank our collaborators with the Communidad Native Matses Anexo San Mateo, Caserío de Tunapuco, Centre MURAZ Research Institute, and the Ministry of Health in Burkina Faso for their collaboration and for opening their communities to our research. We thank Dr. Marielle Hoefnagels and students of the OU BioWriting class for their assistance with editing and reviewing the manuscript.

Supplementary Figures 2.1-2.4

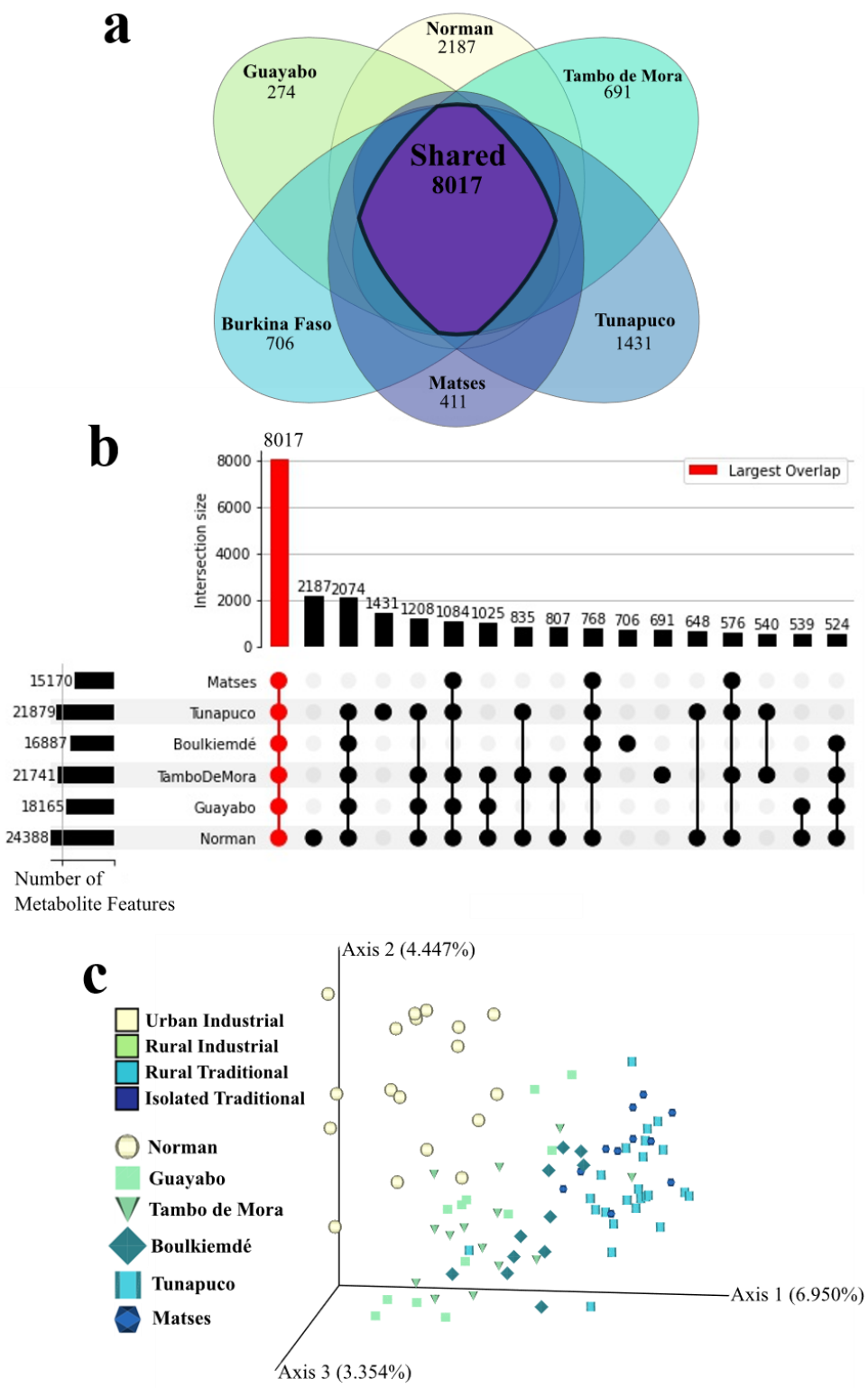
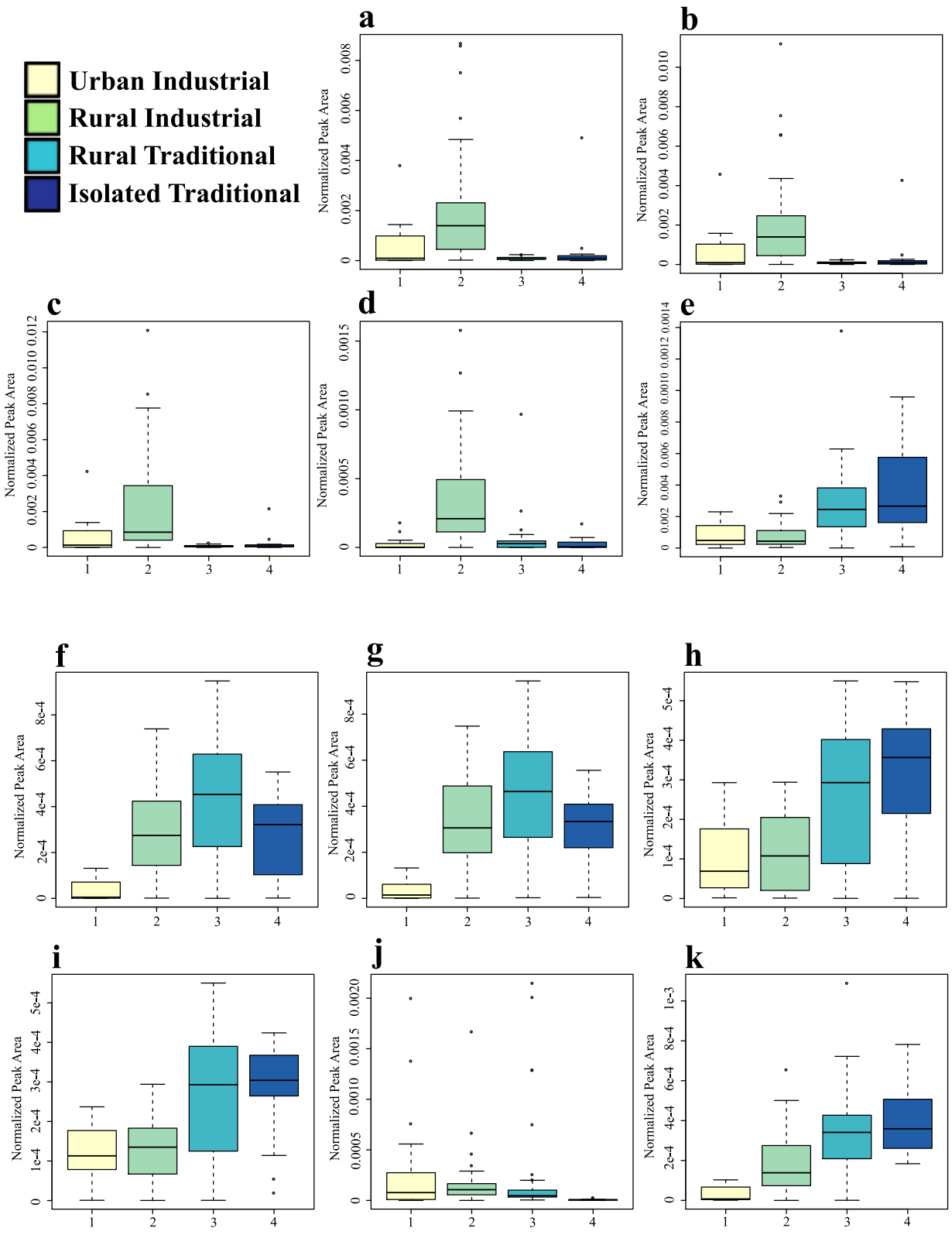
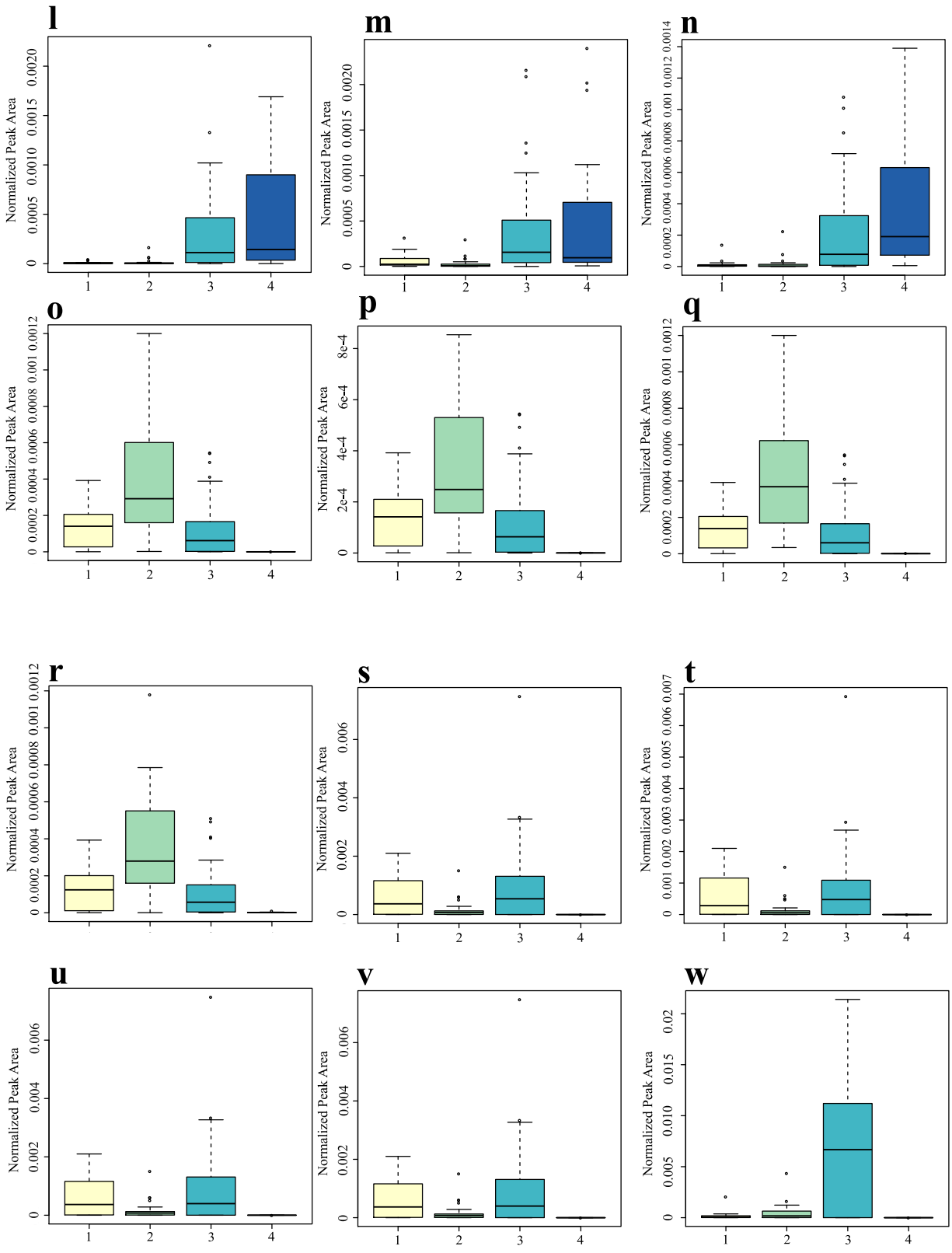
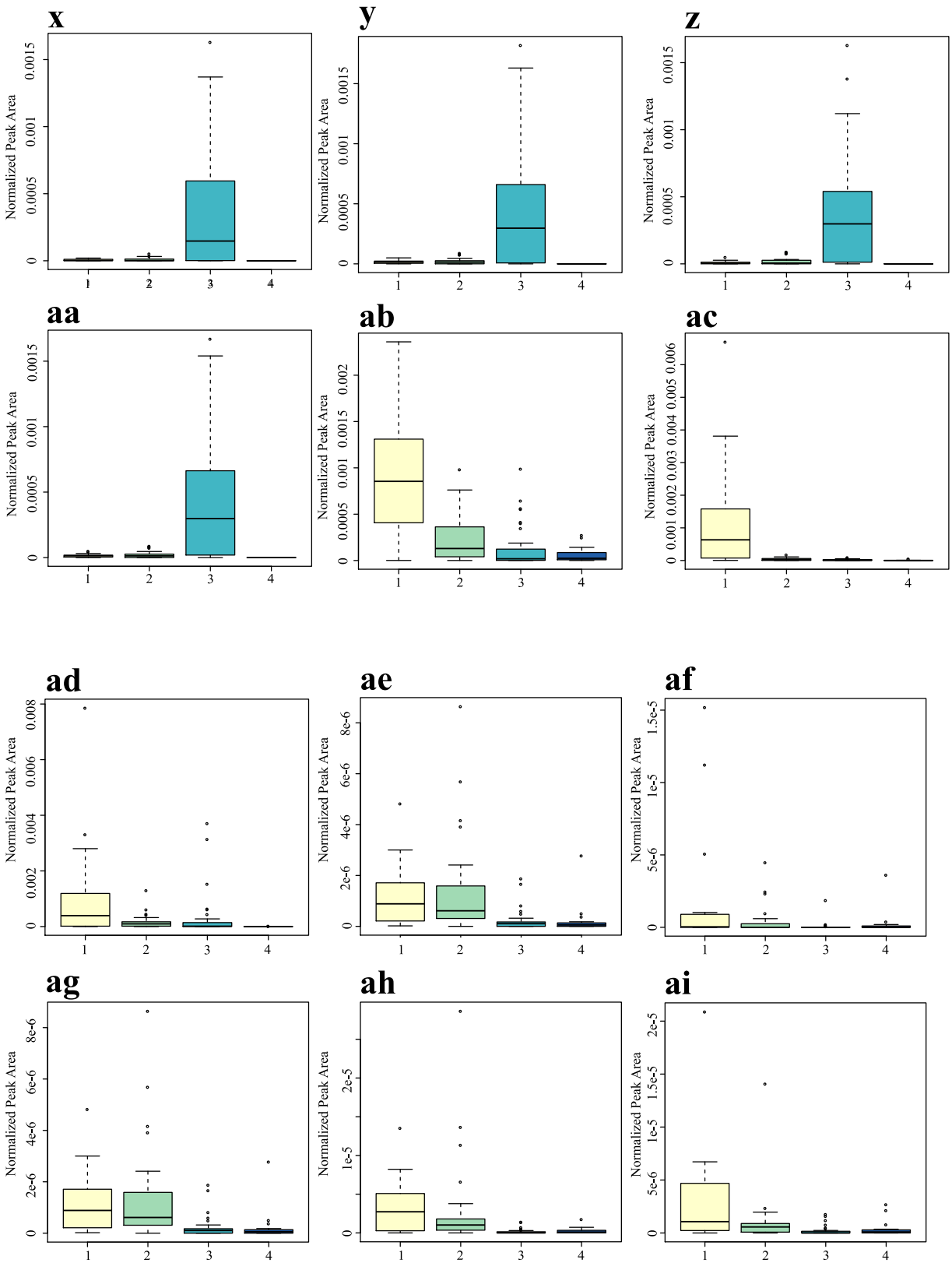


Figure S2.1: Non-gap-filled analyses report an industrialization gradient.
Derived from where n=90 (Norman n=18, Guayabo n=12, Tambo de Mora n=14, Boulkiemdé n=11, Tunapuco n=24, and Matses n=11). ***a****, Metabolic feature overlap across study populations in non-gap-filled data.* ***b****, UpSet plot of non-gap-filled data displays strong overlap between all samples.* ***c****, Principal coordinate analysis plot derived from Canberra distances of non-gap-filled data exhibits an industrialization gradient, indicating this pattern is not an artifact of data processing procedures. Samples are colored by industrialization category and shape-coded by population. Population samples stored in freezer within one day of collection (Norman and Boulkiemdé) are increased in size compared to samples stored within four days of collection.*


Urban Industrial
Rural Industrial
Rural Traditional
Isolated Traditional
a
b
c
d
e
f
g
h
i
j
k
Normalized Peak Area


l
m
n
o
p
q
r
s
t
u
v
w
Normalized Peak Area


x
y
z
aa
ab
ac
ad
ae
af
ag
ah
ai
Normalized Peak Area

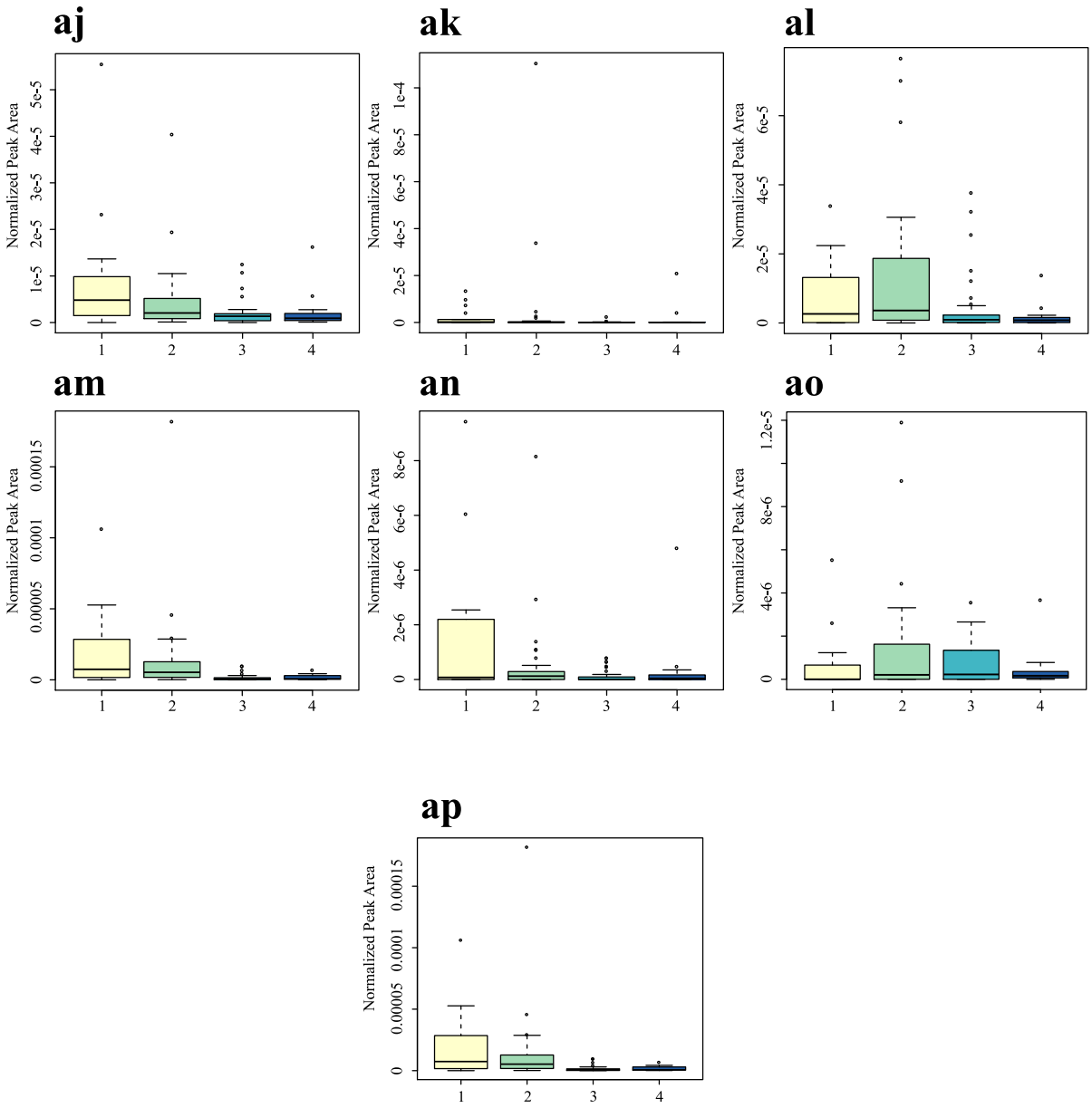
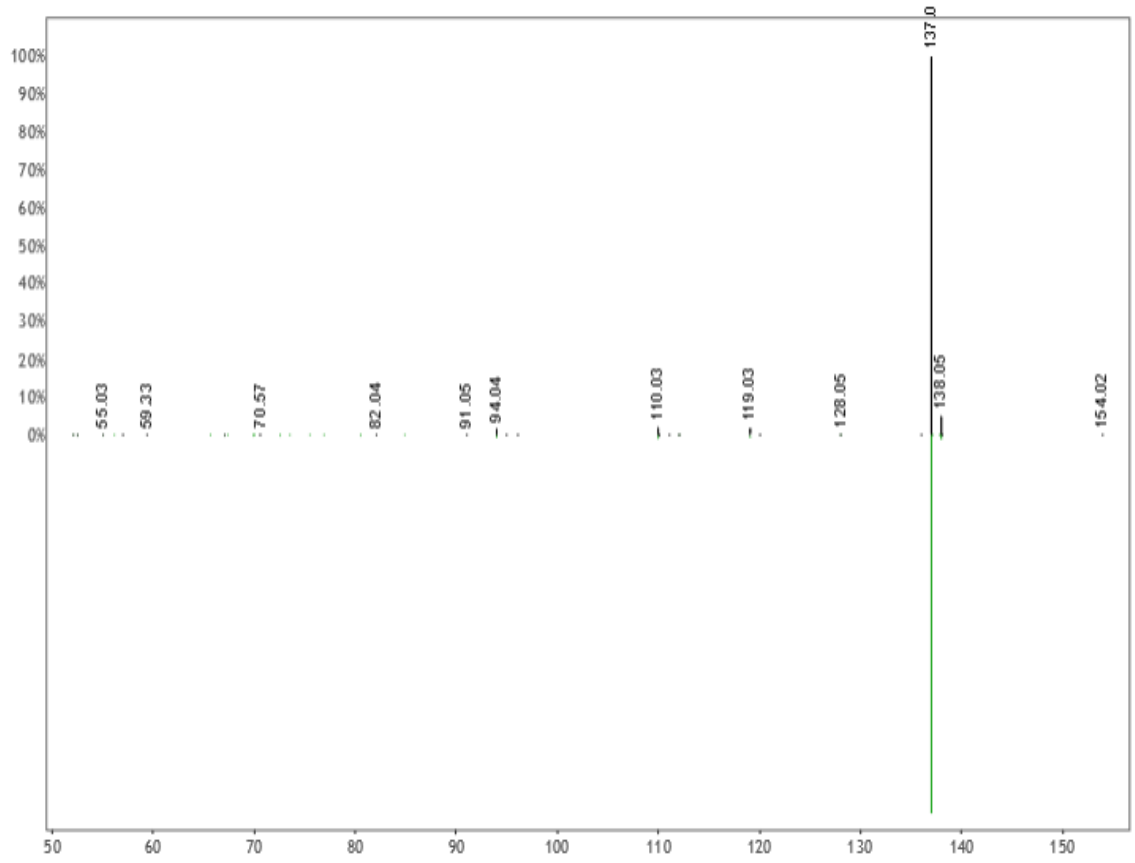

aj
Normalized Peak Area
0
1e-5
2e-5
3e-5
4e-5
5e-5
1
2
3
4
ak
Normalized Peak Area
0
2e-5
4e-5
6e-5
8e-5
1e-4
1
2
3
4
al
Normalized Peak Area
0
2e-5
4e-5
6e-5
1
2
3
4
am
Normalized Peak Area
0
0.00005
0.0001
0.00015
1
2
3
4
an
Normalized Peak Area
0
2e-6
4e-6
6e-6
8e-6
1
2
3
4
ao
Normalized Peak Area
0
4e-6
8e-6
1.2e-5
1
2
3
4
ap
Normalized Peak Area
0
0.00005
0.0001
0.00015
1
2
3
4

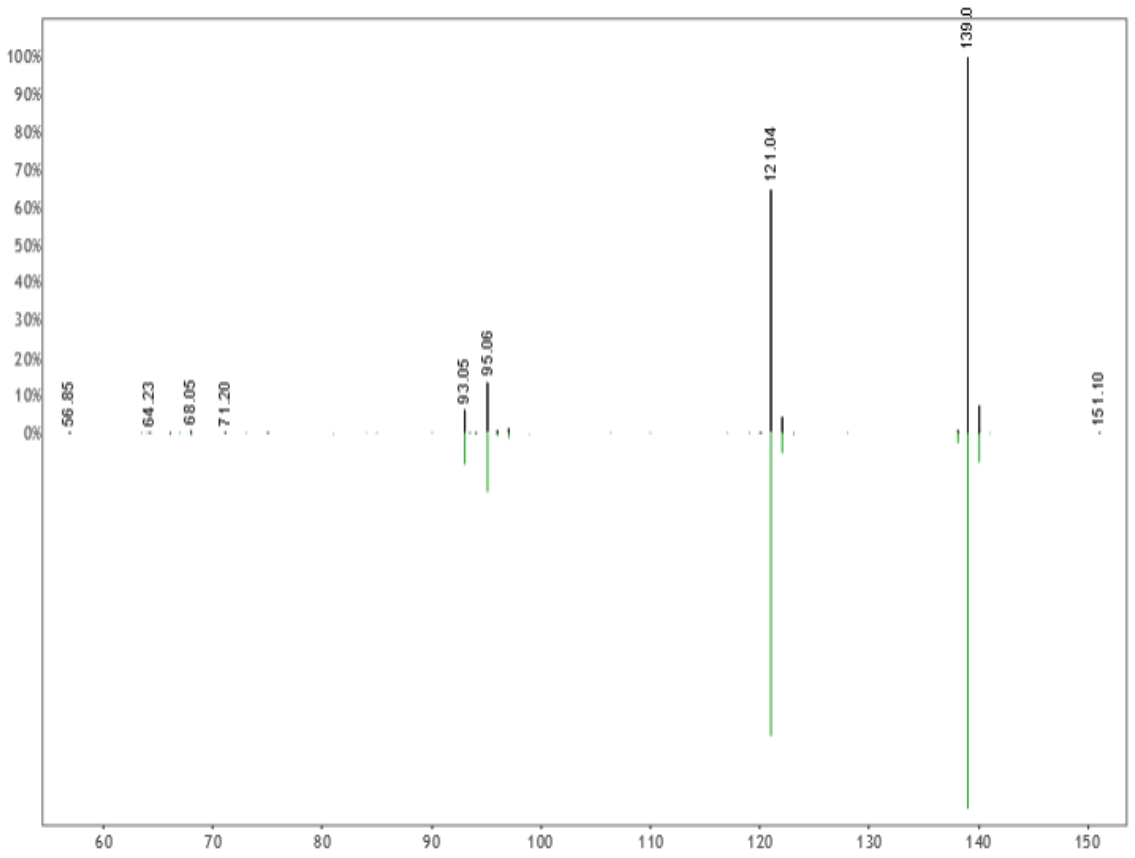
Figure S2.2: Normalized abundances of metabolite features associated with industrialization group differences.

All boxplots are derived from gap-filled analyses. Boxplot axis numbers represent different industrialization groups: 1=urban industrial, 2=rural industrial, 3=rural traditional, 4=isolated traditional. Color key applies for figures a-ap. a-ad, top 30 differential metabolite features based on industrialization group identified by RandomForest. ***a****, m/z 145.134, RT 0.32 min;* ***b****, m/z 145.134, RT 0.32 min;* ***c****, m/z 145.1340, RT 0.36 min;* ***d****, m/z 235.166, RT 0.25 min;* ***e****, m/z 246.186, RT 3.27 min, annotation: leu-leu;* ***f****, m/z 276.108, RT 4.41 min, annotation: N-acetyl-D-glucosamine;* ***g****, m/z 276.108, RT 0.42 min, annotation: N-acetylmuramic acid;* ***h****, m/z 286.176, RT 1.41 min, annotation: ile-pro;* ***i****, m/z 286.176, RT 1.677 min, annotation: phe-pro;* ***j****, m/z 305.186, RT 3.74 min;* ***k****, m/z 332.074, RT 0.36 min;* ***l****, m/z 363.211, RT 1.02;* ***m****, m/z 363.213, RT 0.87 min;* ***n****, m/z 365.192, RT 0.51 min;* ***o****, m/z 379.295, RT 4.8 min;* ***p****, m/z 379.296, RT 4.82 min;* ***q****, m/z 379.296, RT 4.8 min;* ***r****, m/z 379.297, RT 4.81 min;* ***s****, m/z 398.342, RT 4.76 min;* ***t****, m/z 398.342, RT 4.82 min;* ***u****, m/z 398.342, RT 4.84 min;* ***v****, m/z 398.345, RT 4.81 min;* ***w****, m/z 400.358, RT 4.83 min, annotation: uvaol;* ***x****, m/z 414.335, RT 4.49 min;* ***y****, m/z 414.337, RT 4.43 min;* ***z****, m/z 414.337, RT 4.38 min;* ***aa****, m/z 414.337, RT 4.43 min;* ***ab****, m/z 591.318, RT 4.52 min, annotation: urobilin;* ***ac****, m/z 593.333, RT 4.98 min;* ***ad****, m/z 597.37, RT 5.31 min.* ***ae-ap****, amino acid-conjugated bile acids by industrialization group.* ***ae****, leucocholic acid (Kruskal-Wallis p=1.69e-7);* ***af****, tyrosocholic acid (Kruskal-Wallis p=7.71e-3);* ***ag****, glutamate-conjugated chenodeoxycholic acid (Kruskal-Wallis p=1.69e-7);* ***ah****, tryptophan-conjugated chenodeoxycholic acid (Kruskal-Wallis p=4.9e-7);* ***ai****, aspartate-conjugated chenodeoxycholic acid (Kruskal-Wallis p=1.13e-5);* ***aj****, histidine-conjugated chenodeoxycholic acid (Kruskal-Wallis p=6.41e-3);* ***ak****, histidine-conjugated cholic acid (Kruskal-Wallis p=0.04);* ***al****, leucine-conjugated chenodeoxycholic acid (Kruskal-Wallis p=0.04);* ***am****, tyrosine-conjugated deoxycholic acid (Kruskal-Wallis p=1.61e-5).* ***an****, aspartate-conjugated cholic acid (Kruskal-Wallis p=0.05);* ***ao****, threonine-conjugated chenodeoxycholic acid (Kruskal-Wallis p=0.4).* ***ap****, Phenylalanocholic acid (Kruskal-Wallis p=1.904e-6).*

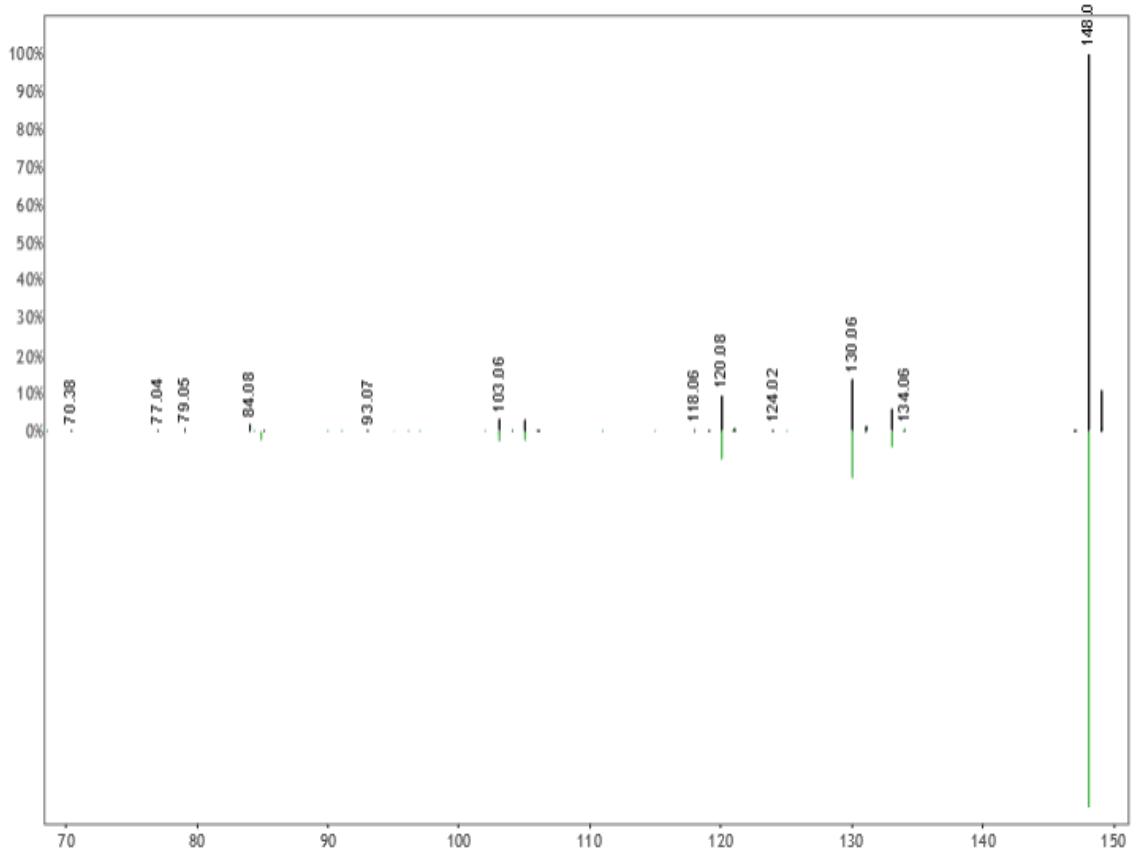
a


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.03
59.33
70.57
82.04
91.05
94.04
110.03
119.03
128.05
137.0
138.05
154.02
50
60
70
80
90
100
110
120
130
140
150

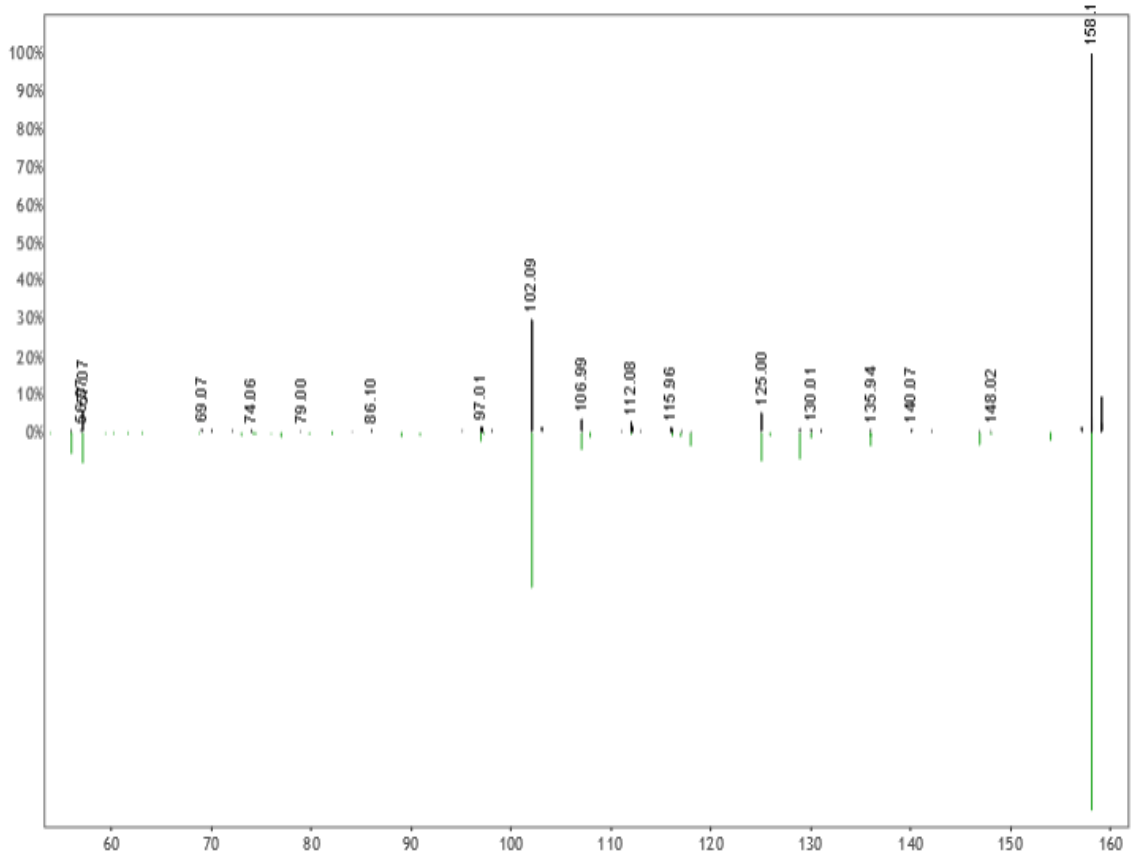
b


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.85
64.23
68.05
71.20
93.05
95.06
121.04
139.0
151.10
60
70
80
90
100
110
120
130
140
150

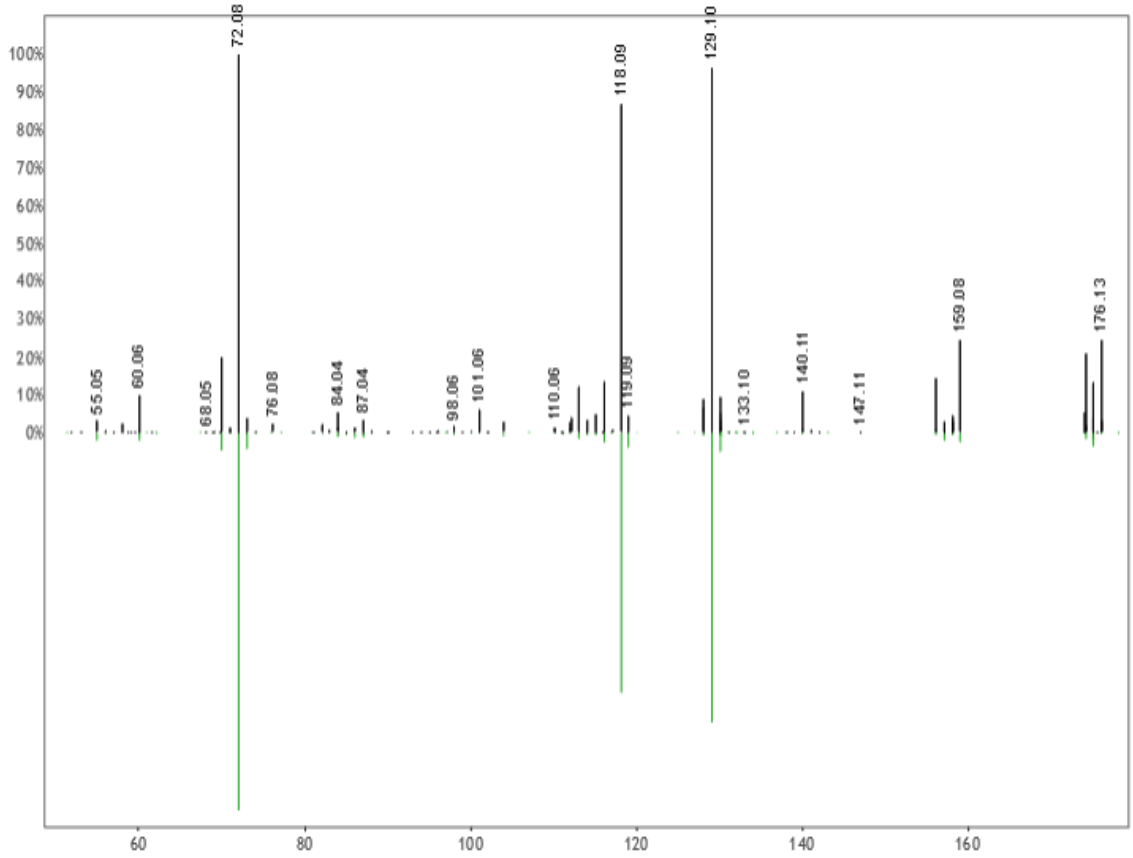
c


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
70.38
77.04
79.05
84.08
93.07
103.06
118.06
120.08
124.02
130.06
134.06
148.0
70
80
90
100
110
120
130
140
150

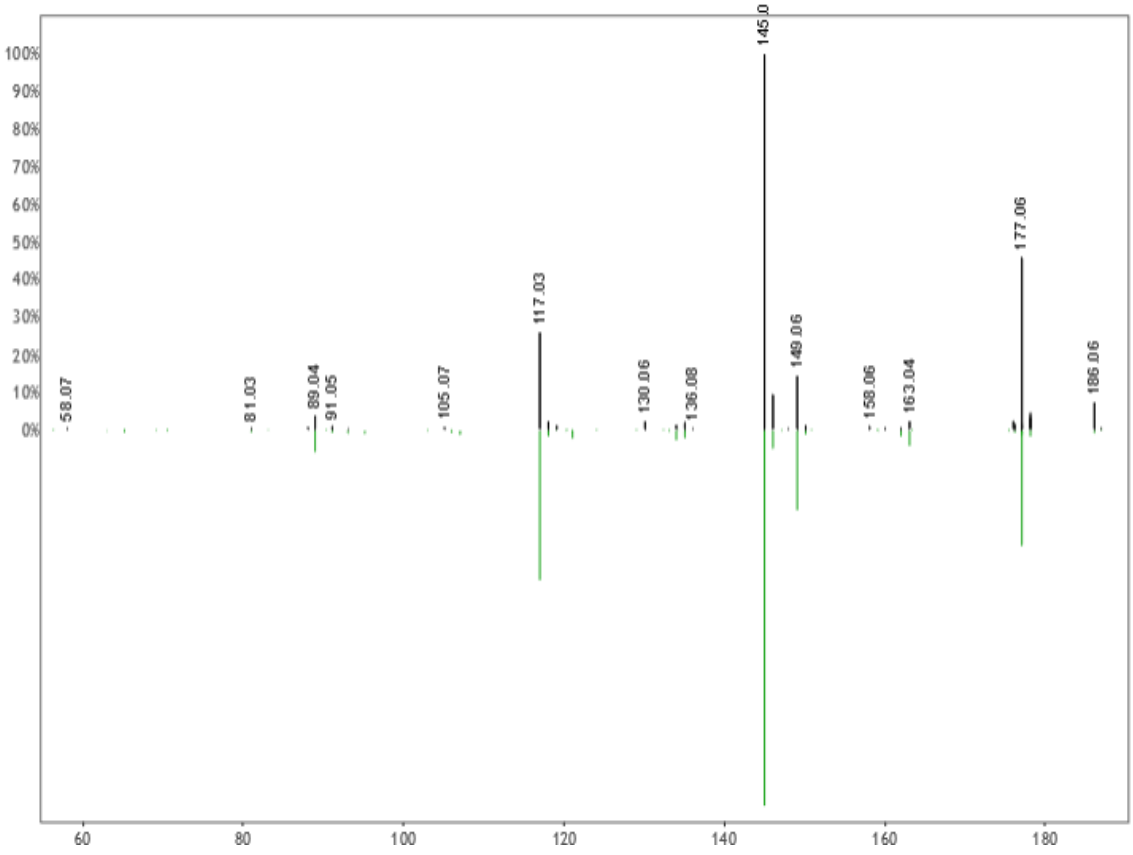
d


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
69.07
74.06
79.00
86.10
97.01
102.09
106.99
112.08
115.96
125.00
130.01
135.94
140.07
148.02
158.1
60
70
80
90
100
110
120
130
140
150
160

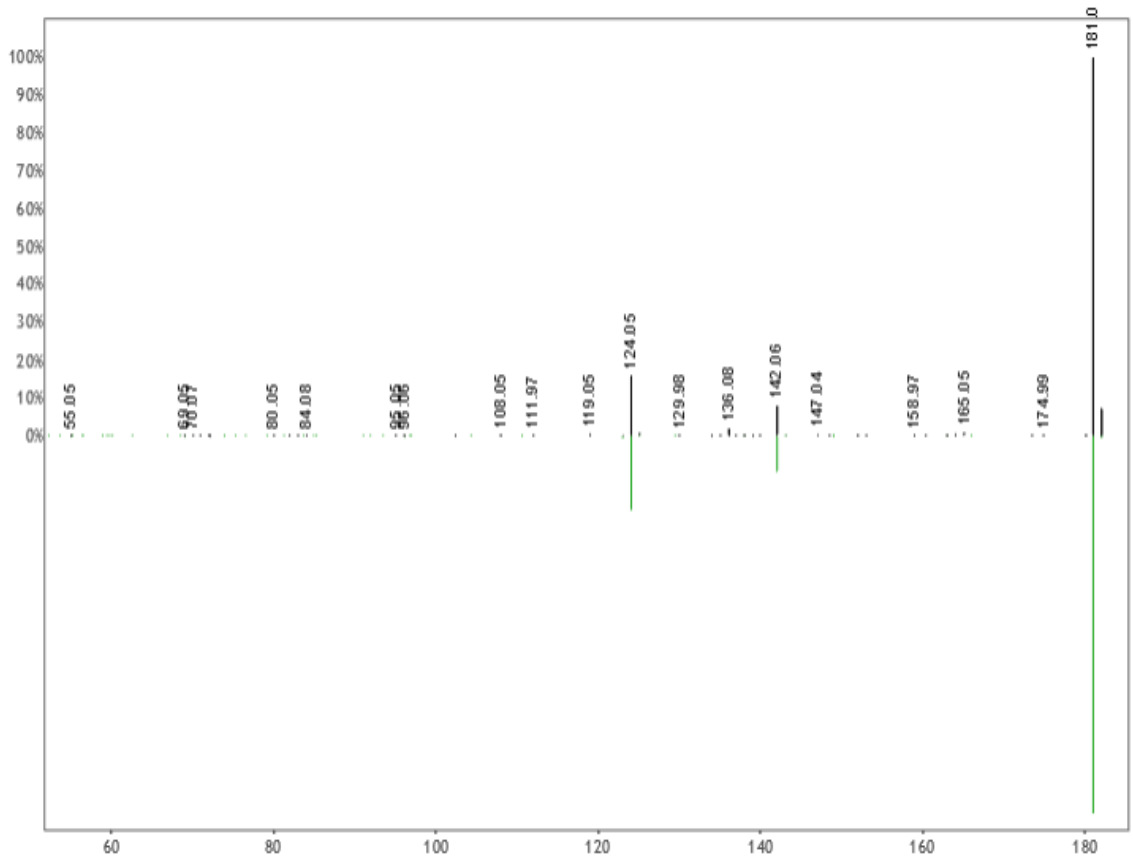
e


72.08
118.09
129.10
55.05
60.06
68.05
76.08
84.04
87.04
98.06
101.06
110.06
119.09
133.10
140.11
147.11
159.08
176.13
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
60
80
100
120
140
160

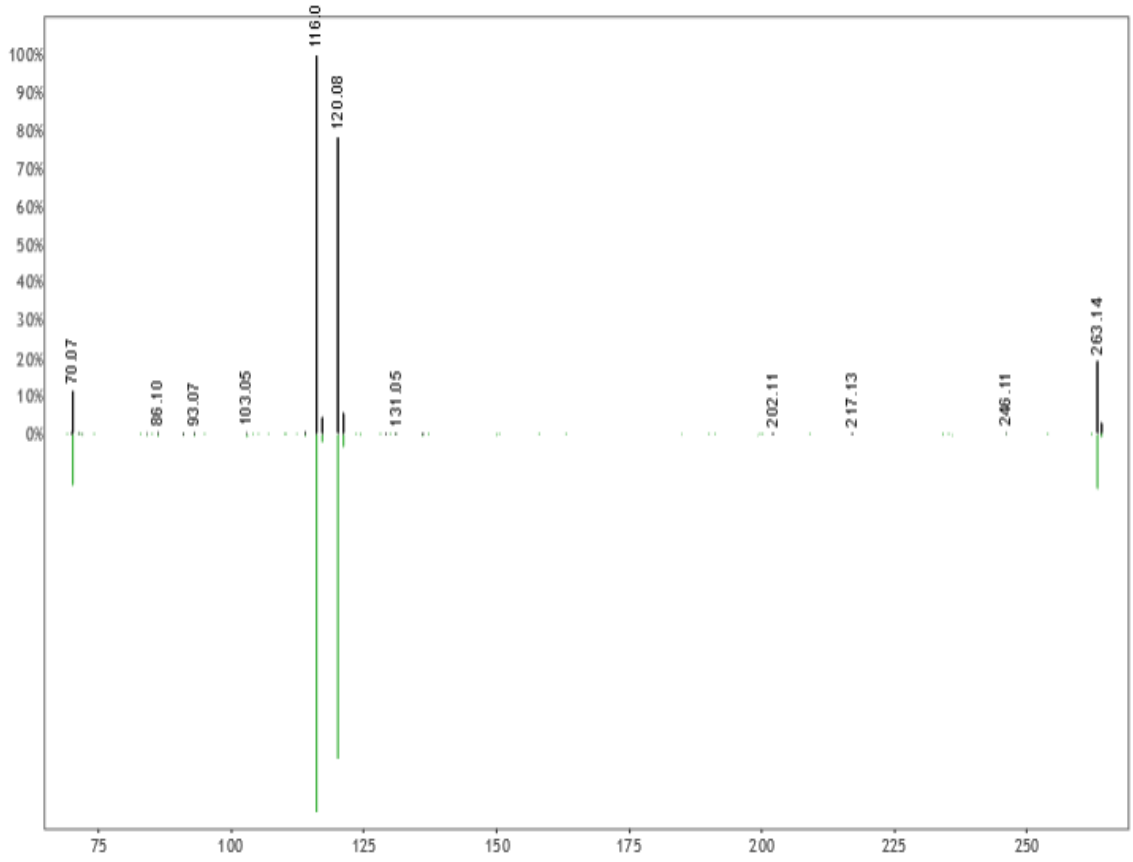
f


145.0
117.03
177.06
149.06
58.07
81.03
89.04
91.05
105.07
130.06
136.08
158.06
163.04
186.06
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
60
80
100
120
140
160
180

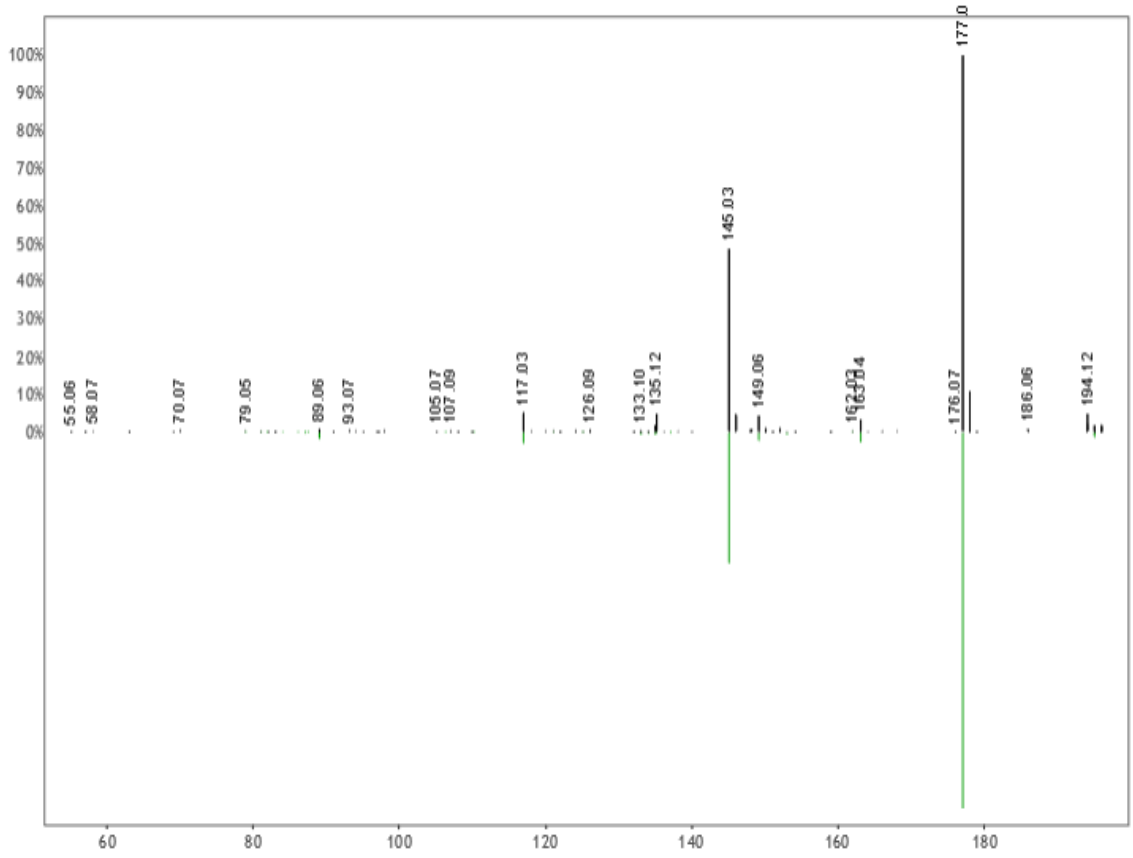
g


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
69.05
70.07
80.05
84.08
96.06
96.96
108.05
111.97
119.05
124.05
129.98
136.08
142.06
147.04
158.97
165.05
174.99
181.0
60
80
100
120
140
160
180

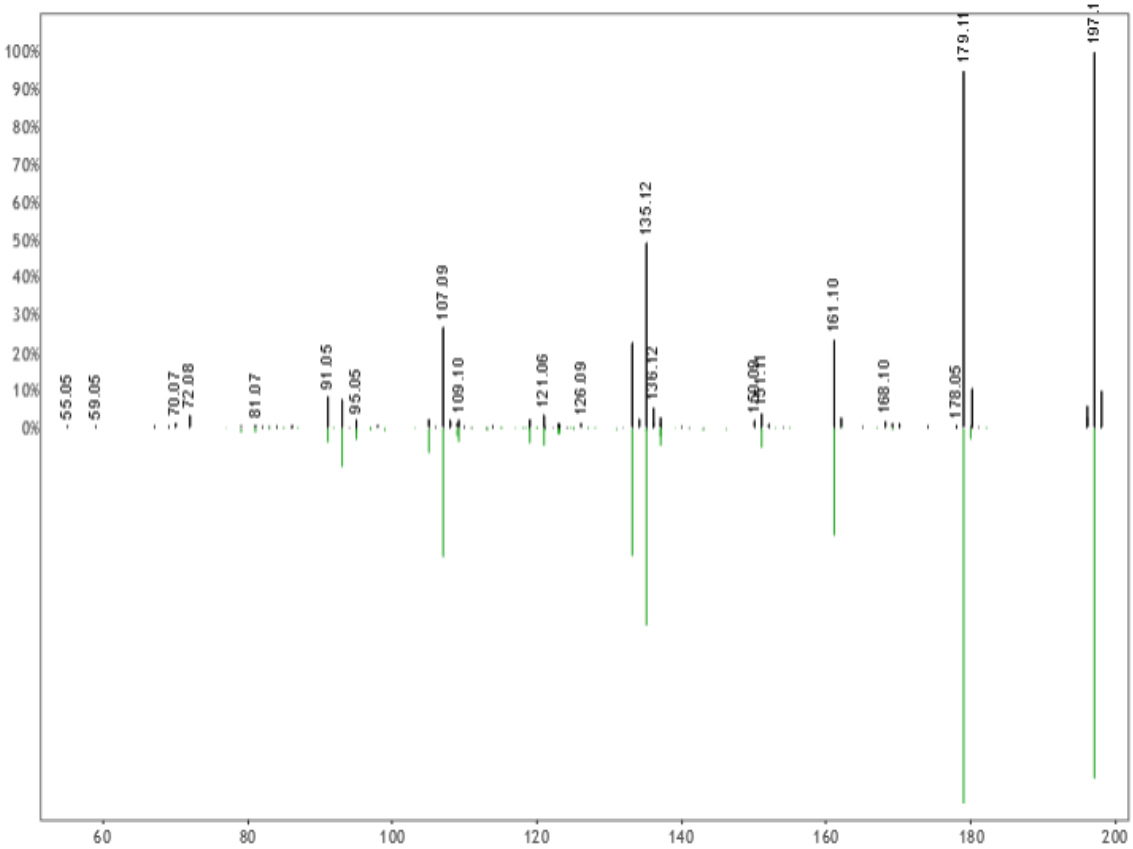
h


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
70.07
86.10
93.07
103.05
116.0
120.08
131.05
202.11
217.13
246.11
263.14
75
100
125
150
175
200
225
250

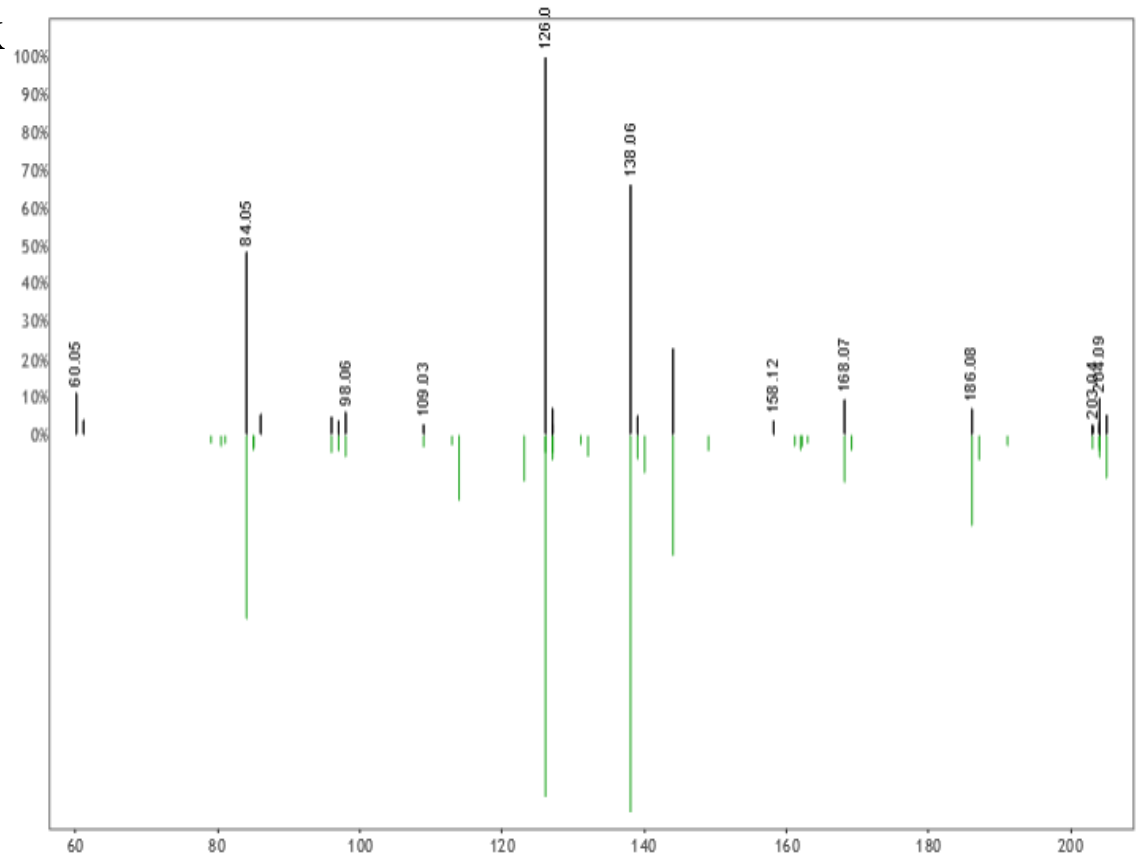
i


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.06
58.07
70.07
79.05
89.06
93.07
105.07
107.09
117.03
126.09
133.10
135.12
145.03
149.06
163.04
176.07
177.0
186.06
194.12
60
80
100
120
140
160
180

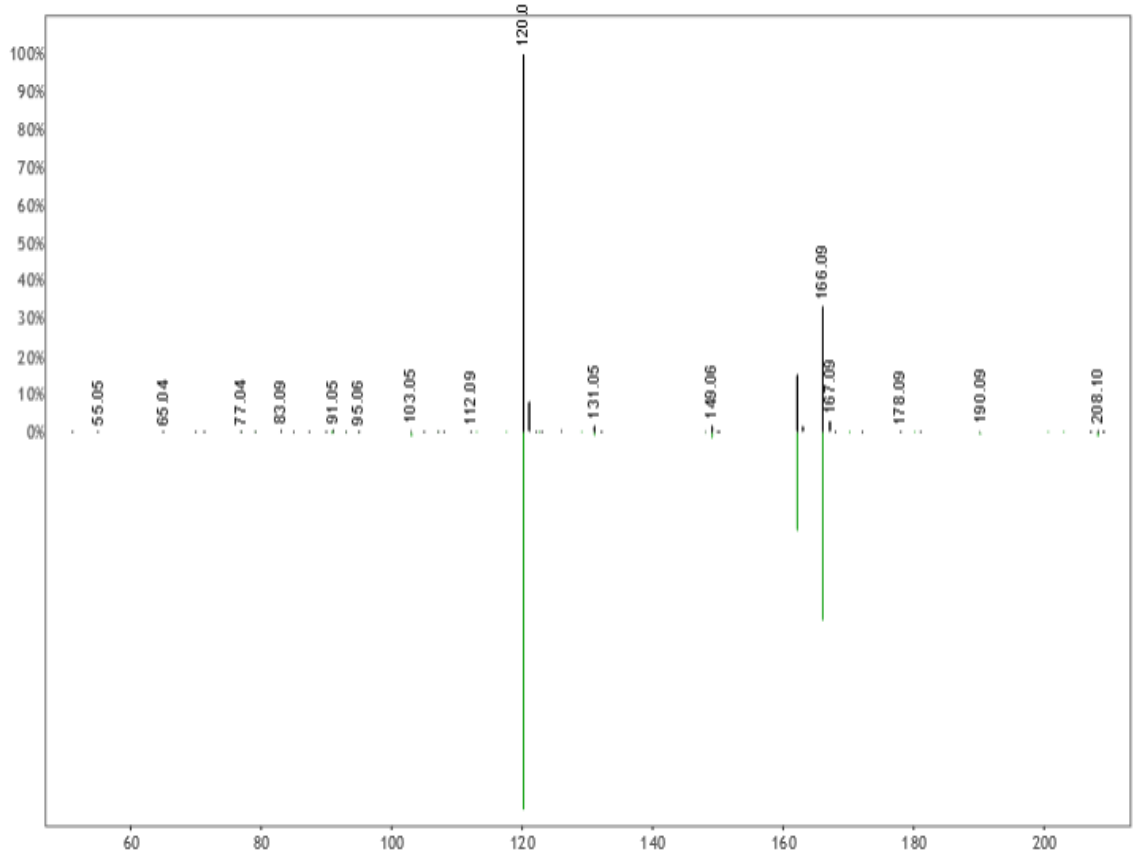
j


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
59.05
70.07
72.08
81.07
91.05
95.05
107.09
109.10
121.06
126.09
135.12
136.12
161.10
168.10
178.05
179.11
197.1
60
80
100
120
140
160
180
200

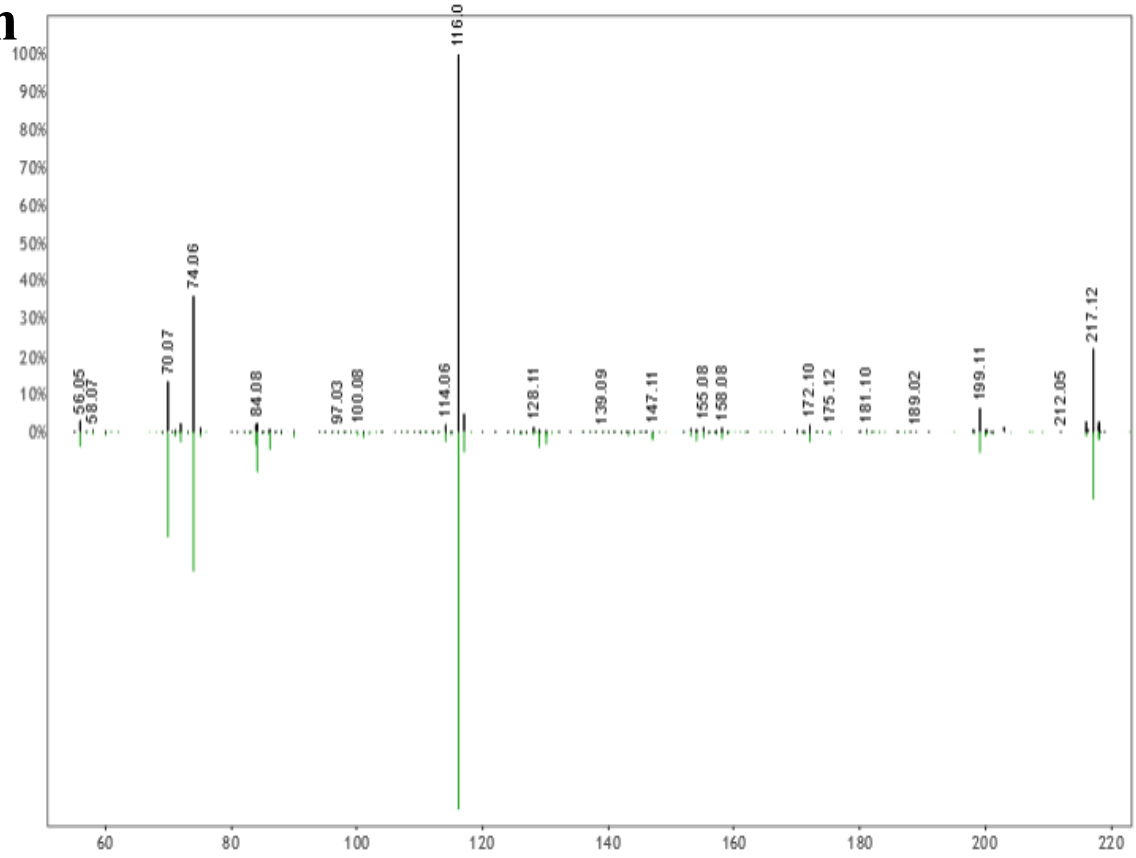
k


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
60.05
84.05
98.06
109.03
126.0
138.06
158.12
168.07
186.08
60
80
100
120
140
160
180
200

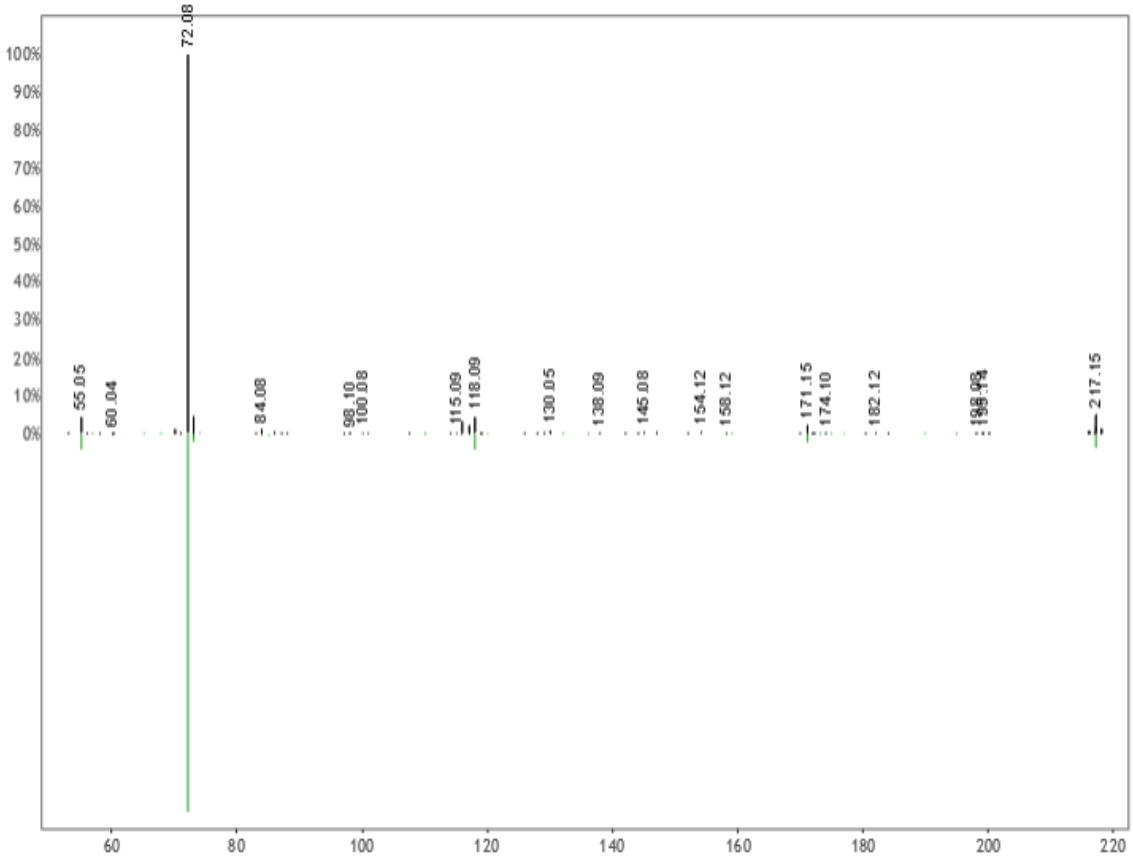
l


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
65.04
77.04
83.09
91.05
95.06
103.05
112.09
120.0
131.05
149.06
166.09
167.09
178.09
190.09
208.10
60
80
100
120
140
160
180
200

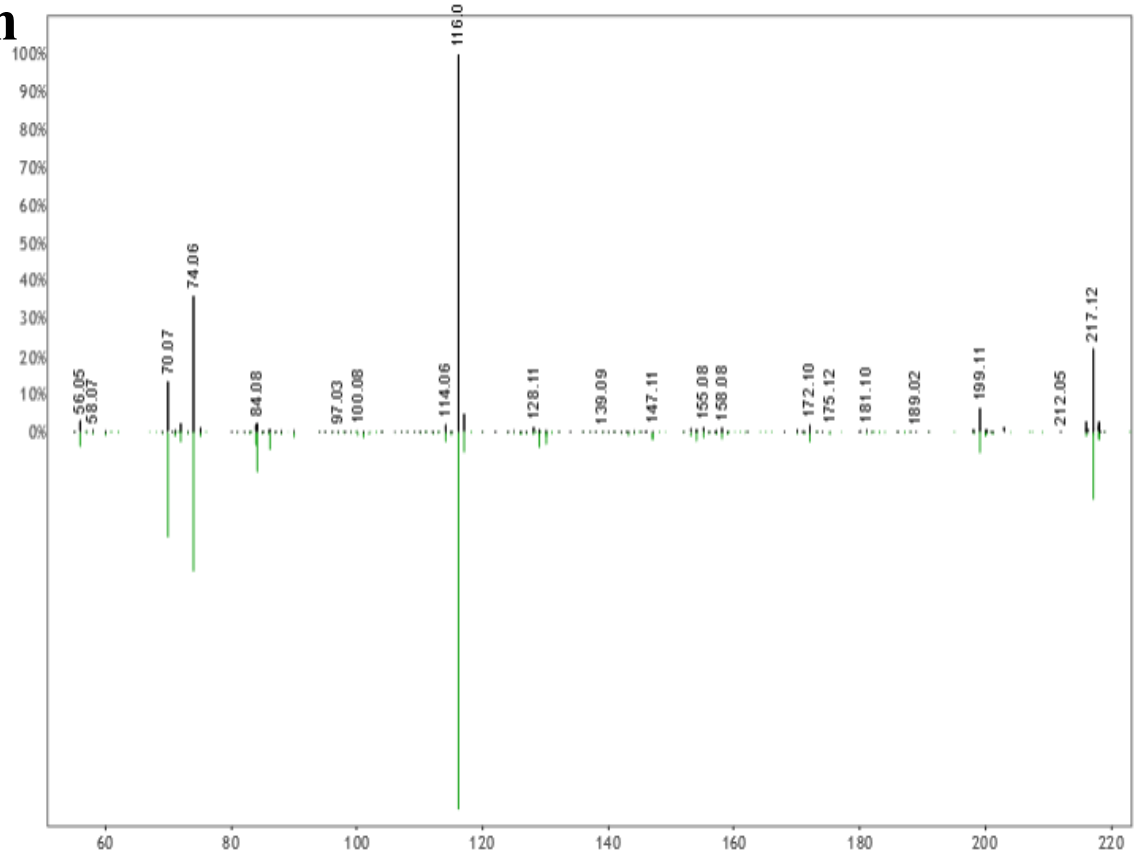
m


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
58.07
70.07
74.06
84.08
97.03
100.08
114.06
116.0
128.11
139.09
147.11
155.08
158.08
172.10
175.12
181.10
189.02
199.11
212.05
217.12
60
80
100
120
140
160
180
200
220

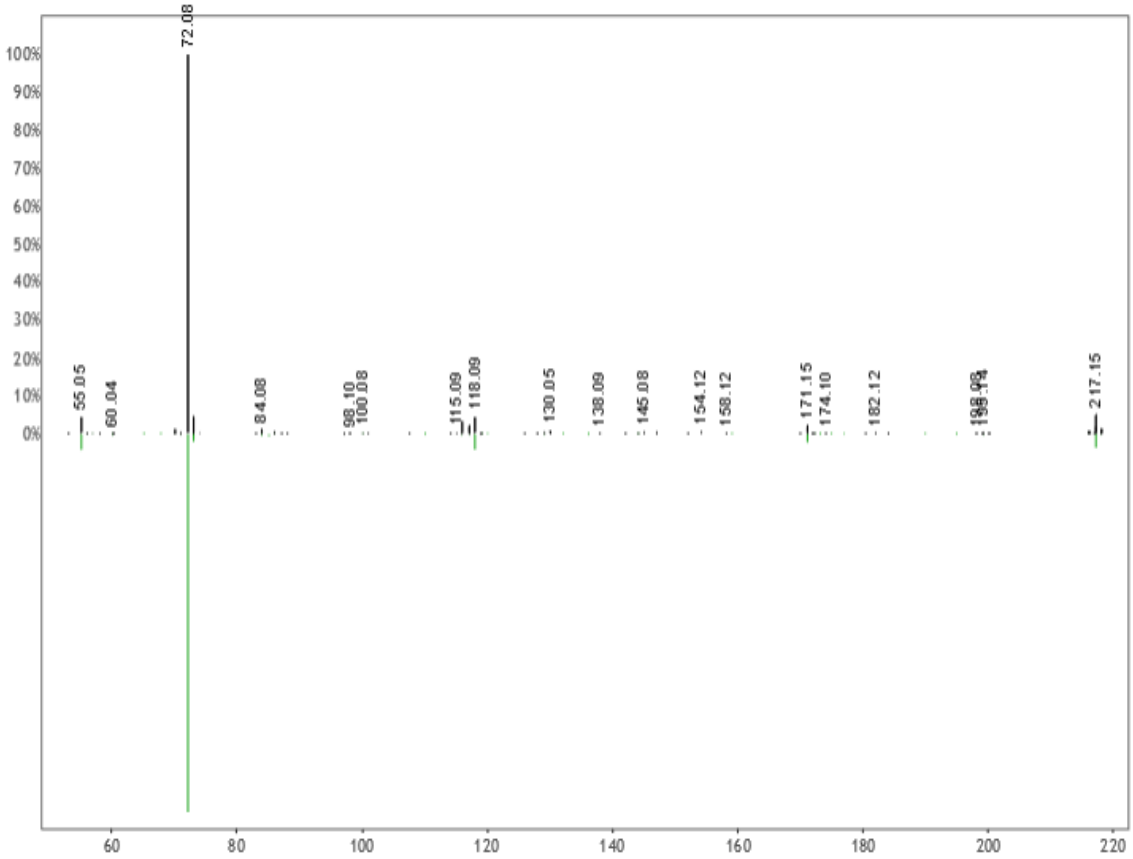
n


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
60.04
72.08
84.08
98.10
100.08
115.09
118.09
130.05
138.09
145.08
154.12
158.12
171.15
174.10
182.12
217.15
60
80
100
120
140
160
180
200
220

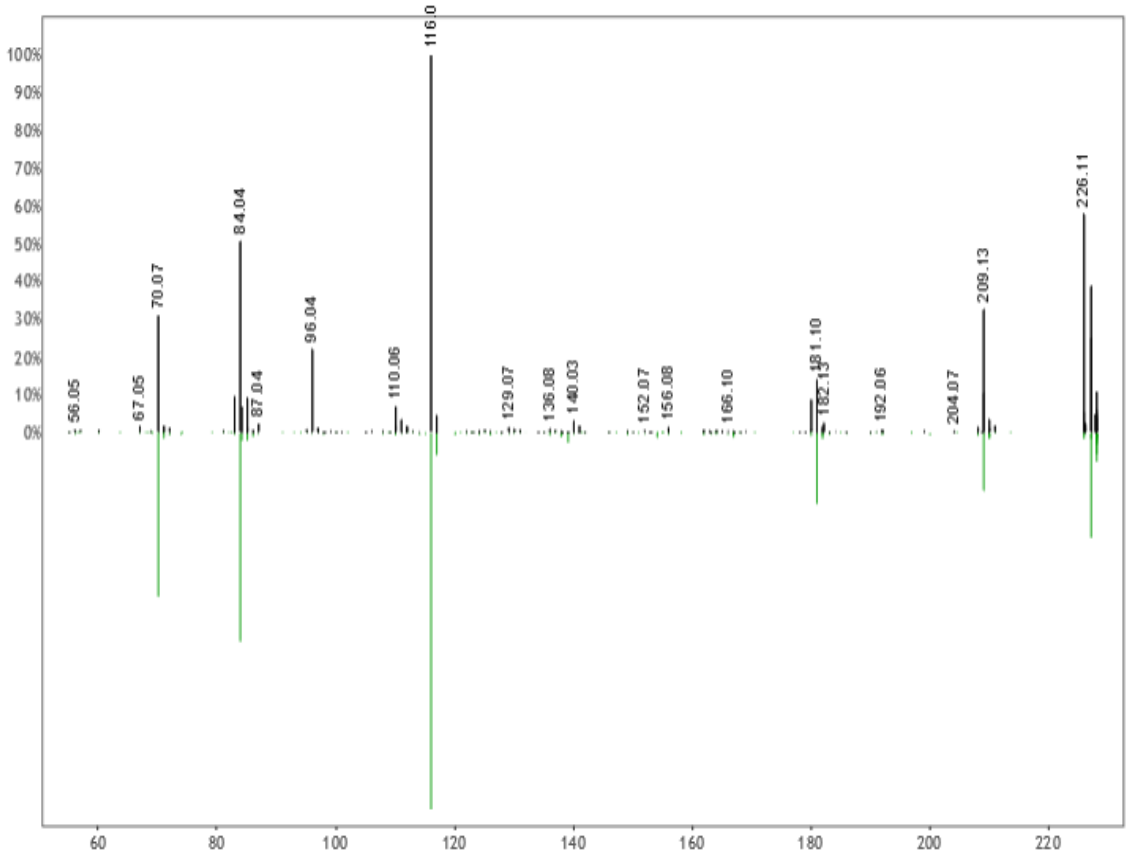
m


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
58.07
70.07
74.06
84.08
97.03
100.08
114.06
116.0
128.11
139.09
147.11
155.08
158.08
172.10
175.12
181.10
189.02
199.11
212.05
217.12
60
80
100
120
140
160
180
200
220

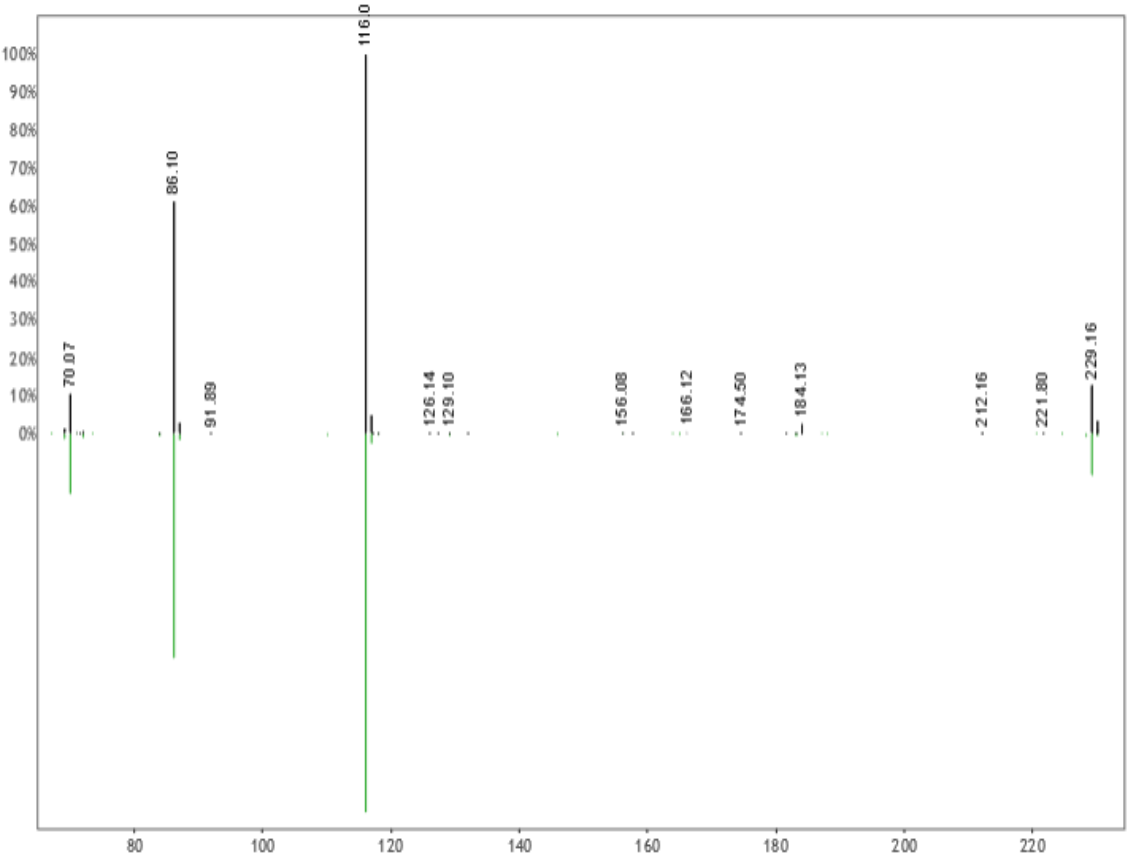
n


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
60.04
72.08
84.08
98.10
100.08
115.09
118.09
130.05
138.09
145.08
154.12
158.12
171.15
174.10
182.12
217.15
60
80
100
120
140
160
180
200
220

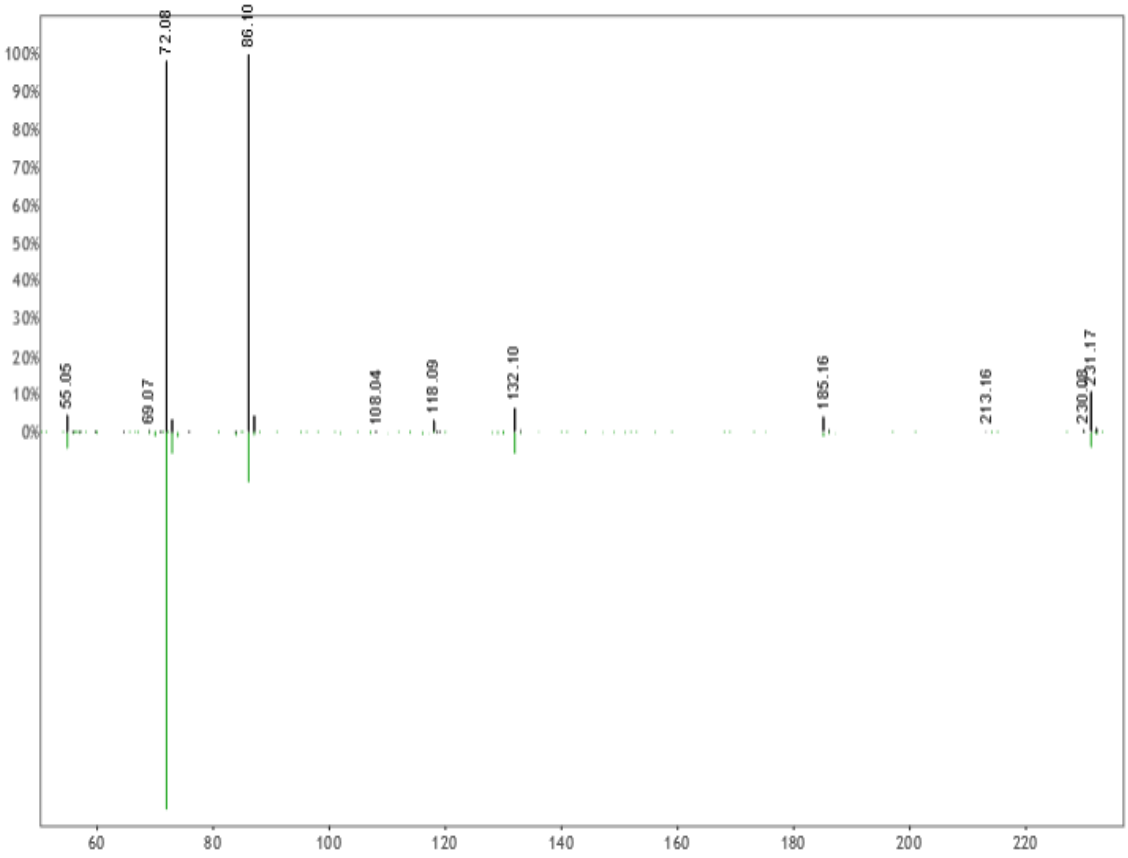
q


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
67.05
70.07
84.04
87.04
96.04
110.06
116.0
129.07
136.08
140.03
152.07
156.08
166.10
181.10
182.13
192.06
204.07
209.13
226.11
60
80
100
120
140
160
180
200
220

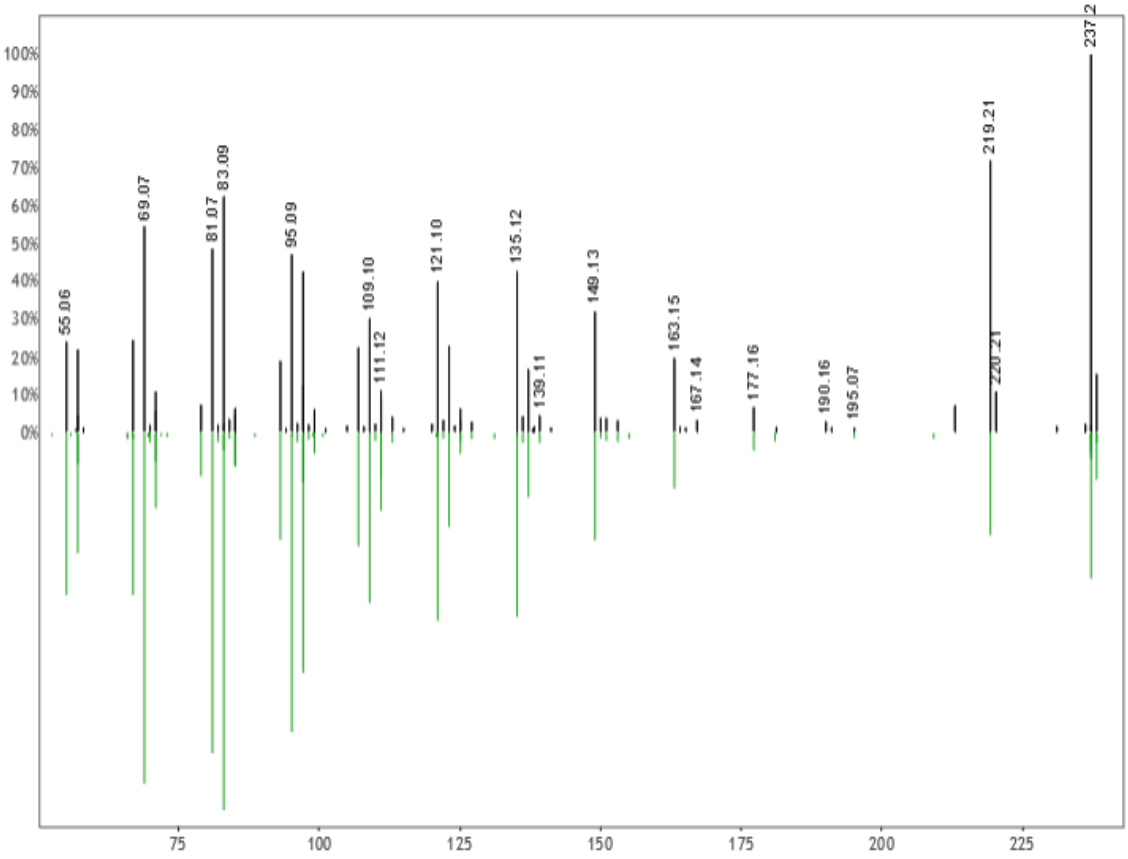
r


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
70.07
86.10
91.89
116.0
126.14
129.10
156.08
166.12
174.50
184.13
212.16
221.80
229.16
80
100
120
140
160
180
200
220

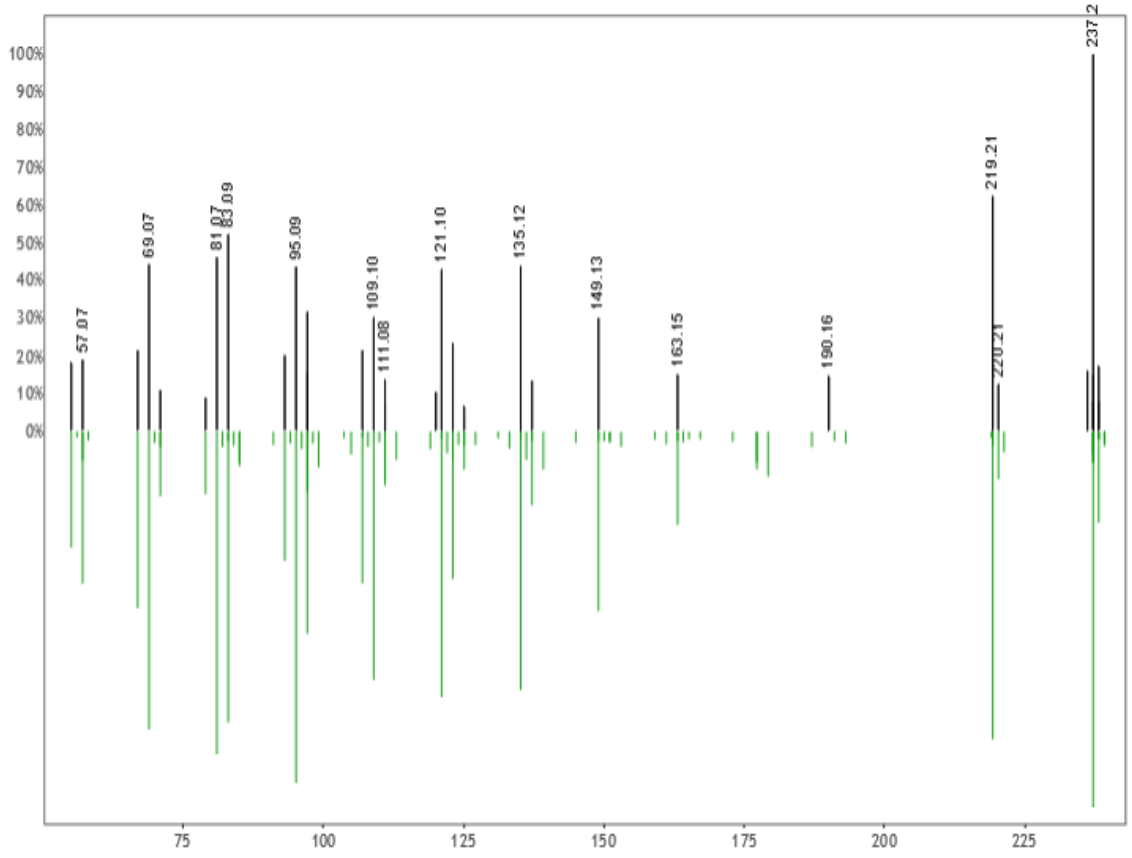
s


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
69.07
72.08
86.10
108.04
118.09
132.10
185.16
213.16
230.08
231.17
60
80
100
120
140
160
180
200
220

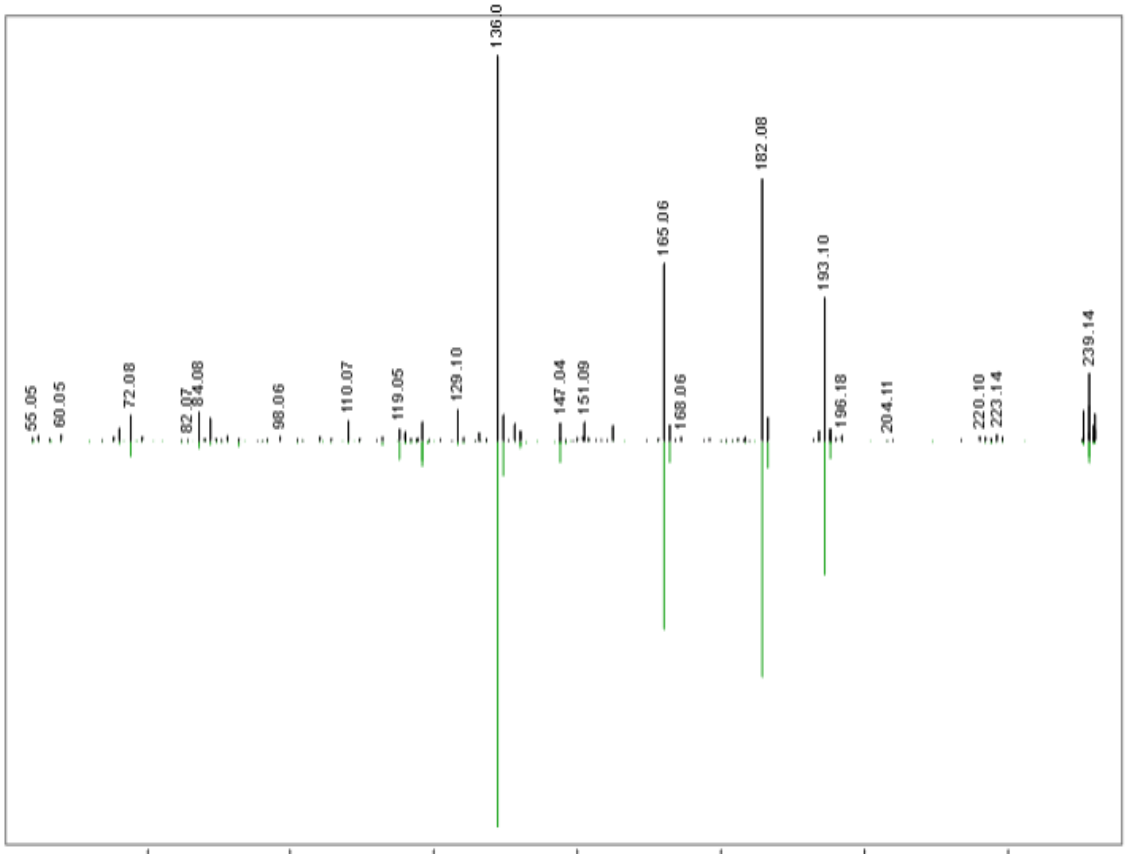
t


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.06
69.07
81.07
83.09
95.09
109.10
111.12
121.10
135.12
139.11
149.13
163.15
167.14
177.16
190.16
195.07
219.21
220.21
237.2
75
100
125
150
175
200
225

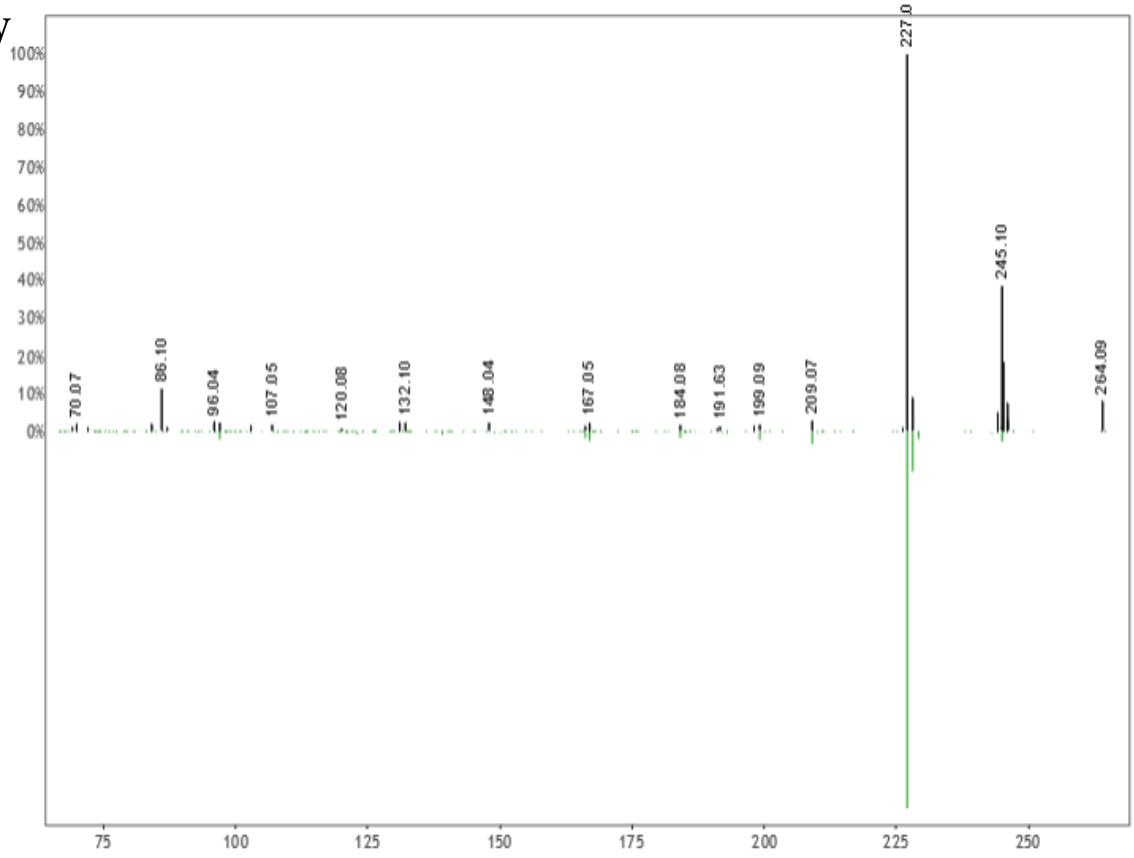
u


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
69.07
81.07
83.09
95.09
109.10
111.08
121.10
135.12
149.13
163.15
190.16
219.21
220.21
237.2
75
100
125
150
175
200
225

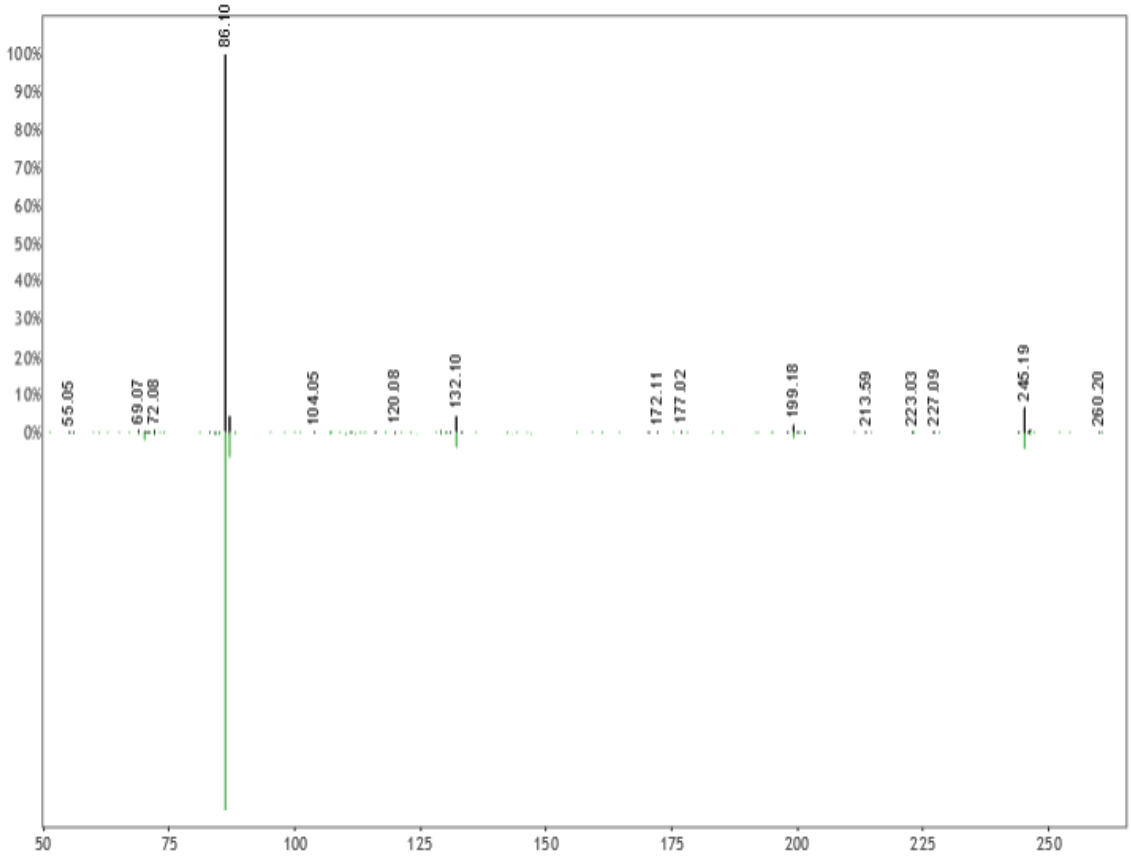
v


55.05
60.05
72.08
82.07
84.08
98.06
110.07
119.05
129.10
136.0
147.04
151.09
165.06
168.06
182.08
193.10
196.18
204.11
220.10
223.14
239.14

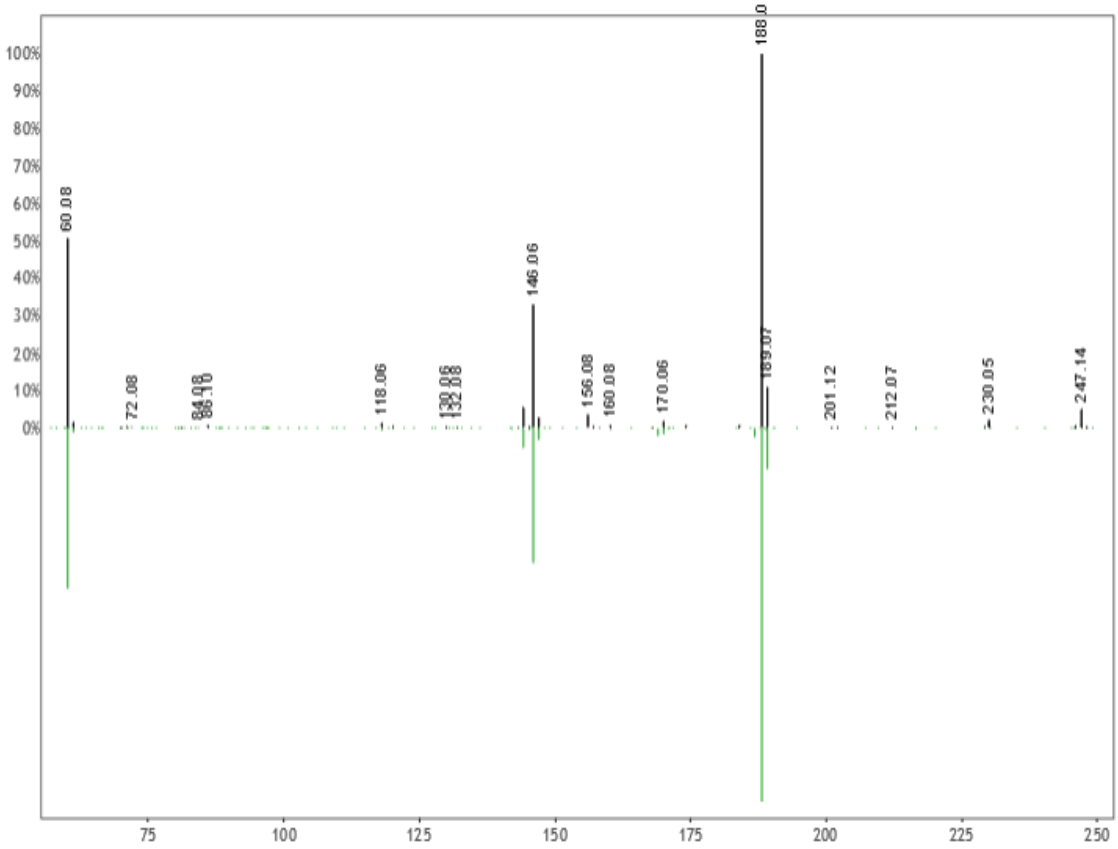
w


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
70.07
86.10
96.04
107.05
120.08
132.10
148.04
167.05
184.08
191.63
199.09
209.07
227.0
245.10
264.09
75
100
125
150
175
200
225
250

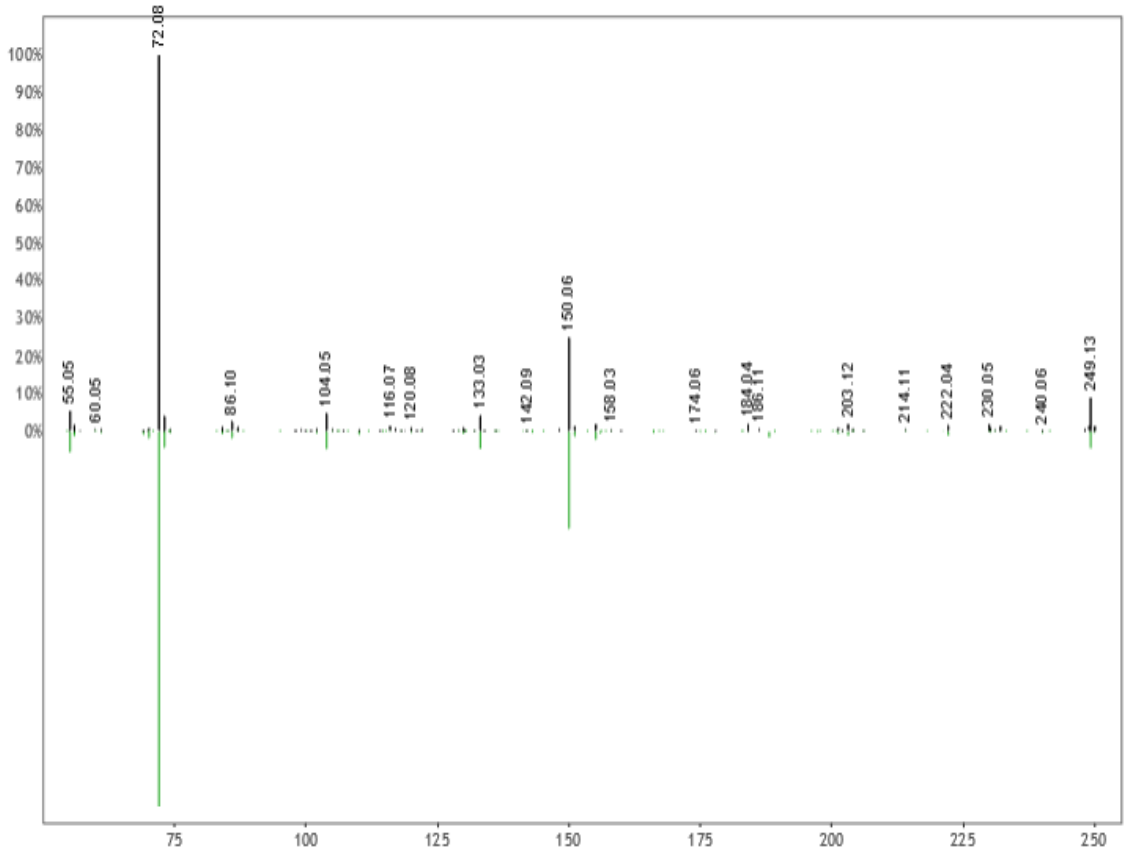
x


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
69.07
72.08
86.10
104.05
120.08
132.10
172.11
177.02
199.18
213.59
223.03
227.09
245.19
260.20
50
75
100
125
150
175
200
225
250

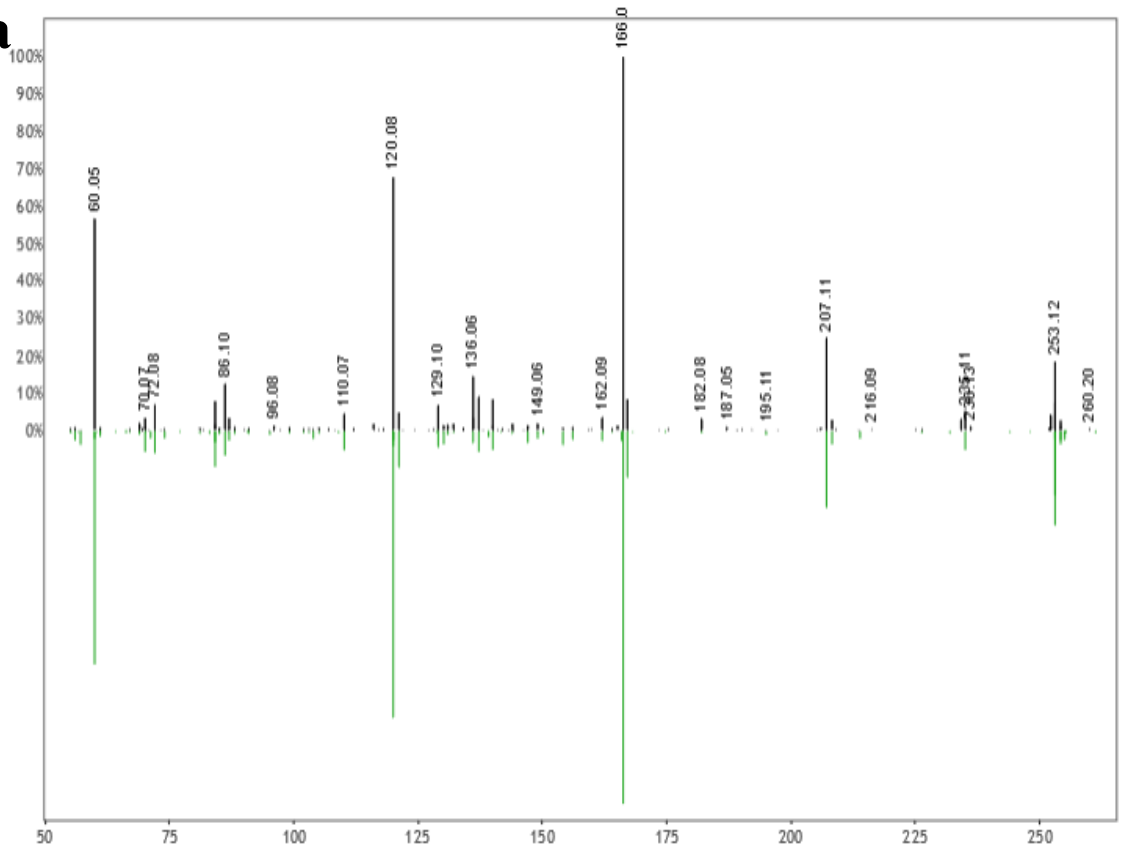
y


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
60.08
72.08
84.08
85.10
118.06
130.06
132.08
146.06
156.08
160.08
170.06
188.0
189.07
201.12
212.07
230.05
247.14
75
100
125
150
175
200
225
250

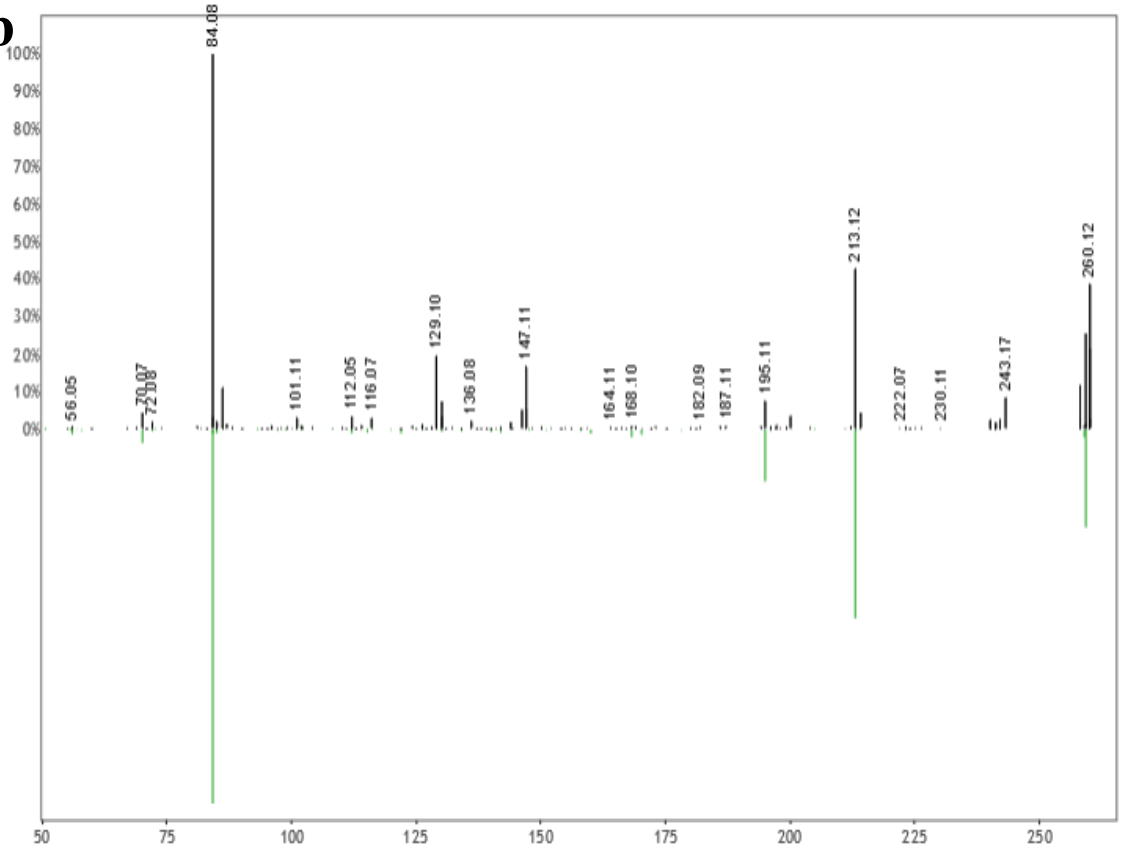
z


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
60.05
72.08
86.10
104.05
116.07
120.08
133.03
142.09
150.06
158.03
174.06
184.04
186.11
203.12
214.11
222.04
230.05
240.06
249.13
75
100
125
150
175
200
225
250

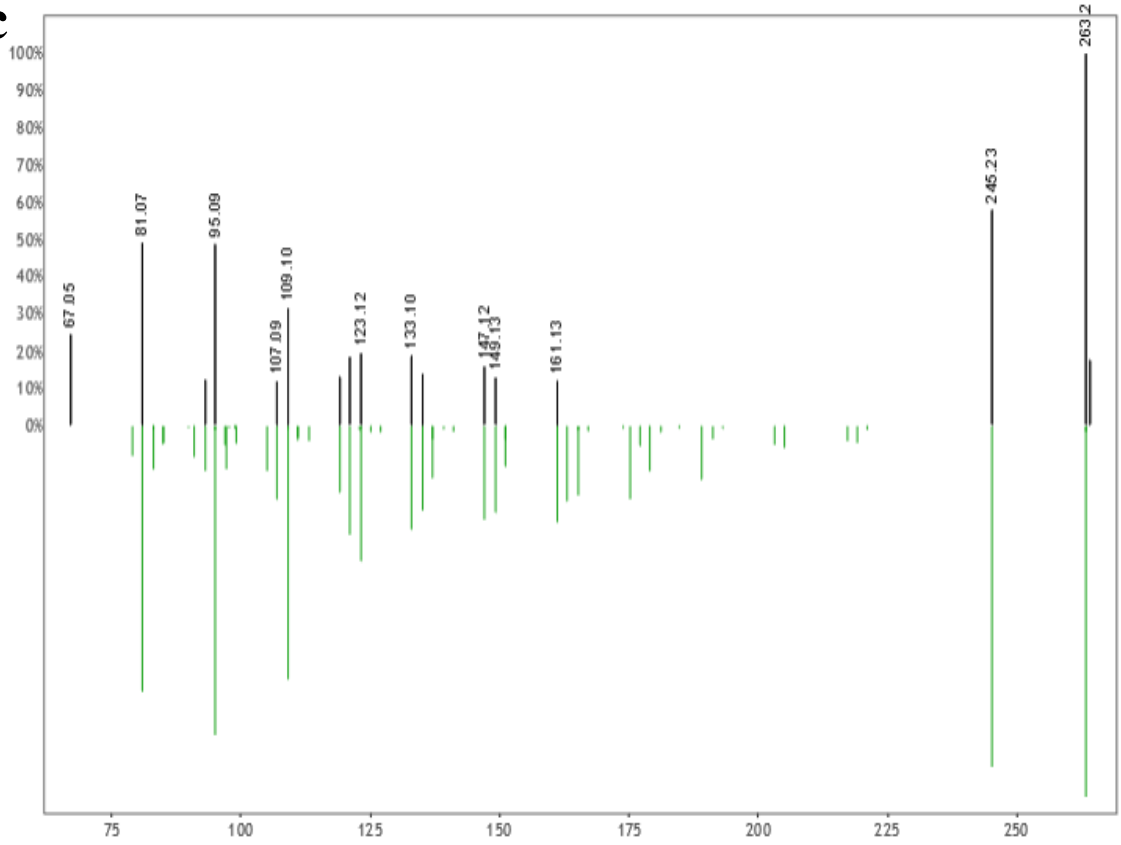
aa


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
60.05
70.07
72.08
86.10
96.08
110.07
120.08
129.10
136.06
149.06
162.09
166.0
182.08
187.05
195.11
207.11
216.09
253.12
260.20
50
75
100
125
150
175
200
225
250

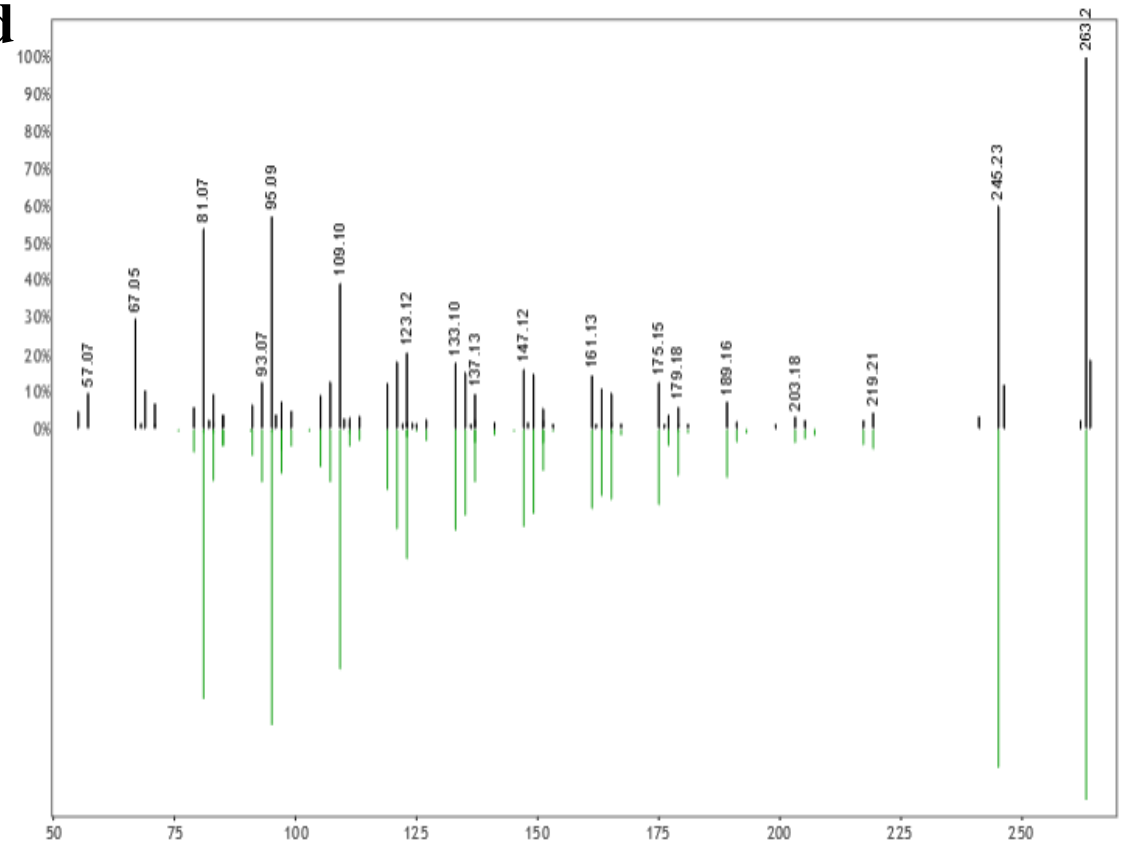
ab


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
70.07
72.08
84.08
101.11
112.05
116.07
129.10
136.08
147.11
164.11
168.10
182.09
187.11
195.11
213.12
222.07
230.11
243.17
260.12
50
75
100
125
150
175
200
225
250

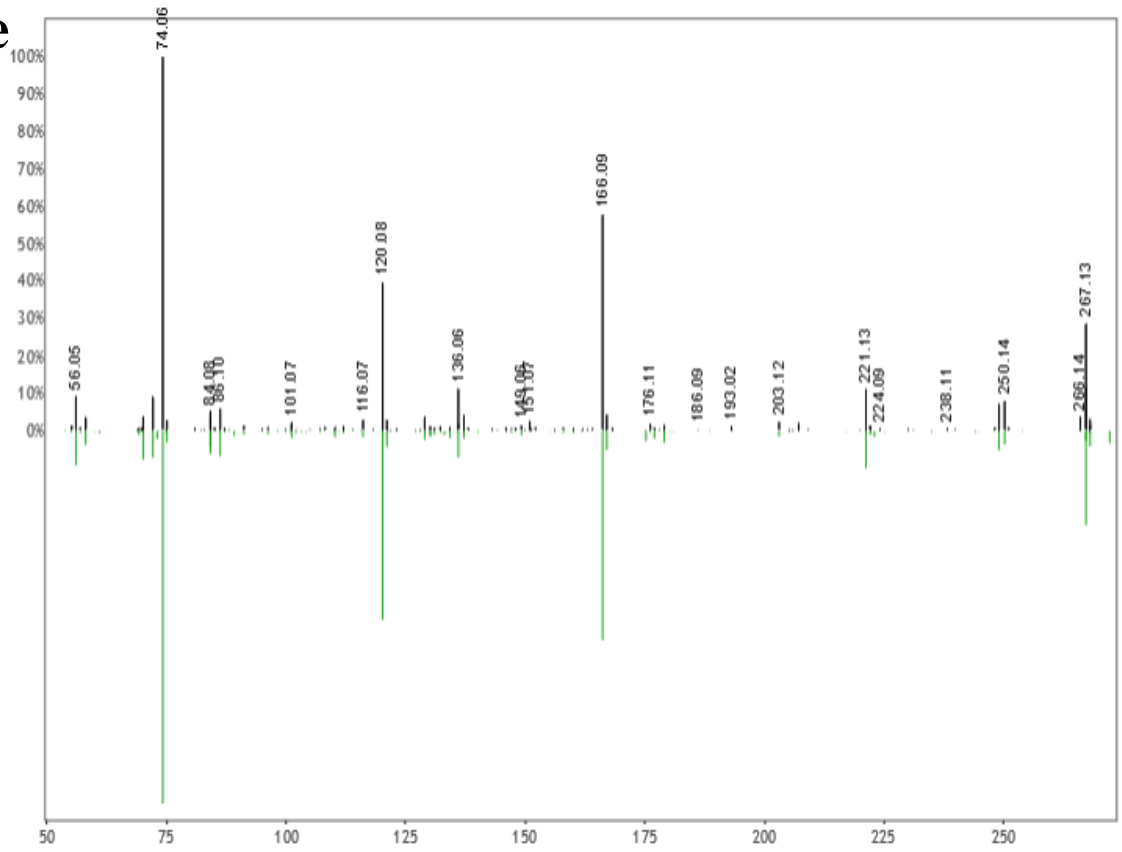
ac


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
67.05
81.07
95.09
107.09
109.10
123.12
133.10
147.12
149.13
161.13
245.23
263.2
75
100
125
150
175
200
225
250

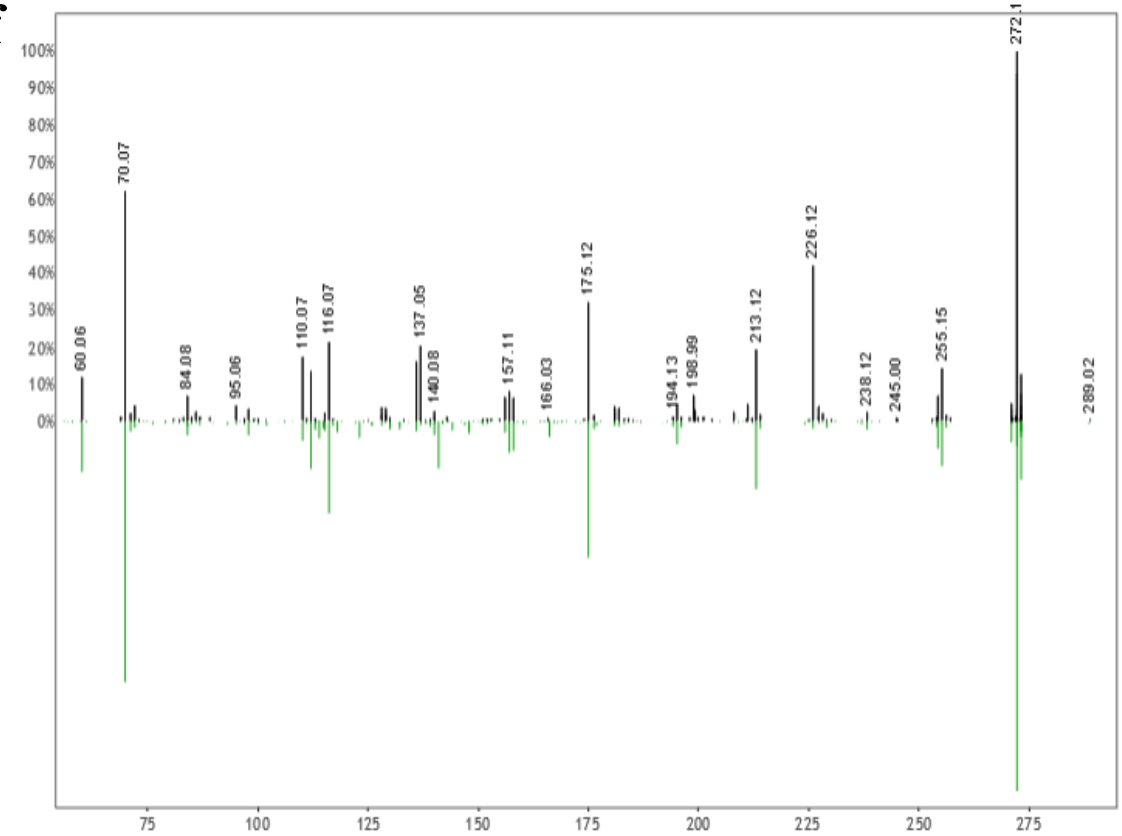
ad


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
67.05
81.07
93.07
95.09
109.10
123.12
133.10
137.13
147.12
161.13
175.15
179.18
189.16
203.18
219.21
245.23
263.2
50
75
100
125
150
175
200
225
250

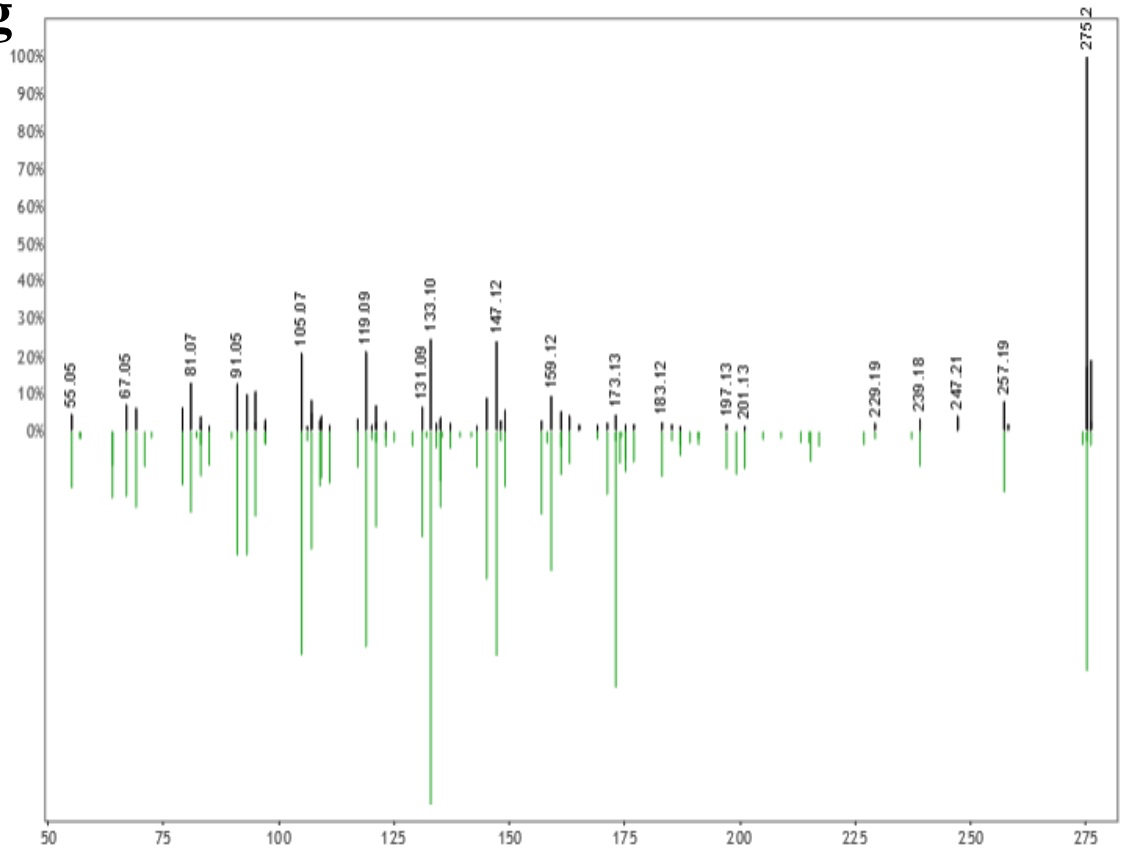
ae


74.06
166.09
120.08
56.05
84.08
88.10
101.07
116.07
136.06
151.07
176.11
186.09
193.02
203.12
221.13
224.09
238.11
250.14
267.13
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
75
100
125
150
175
200
225
250

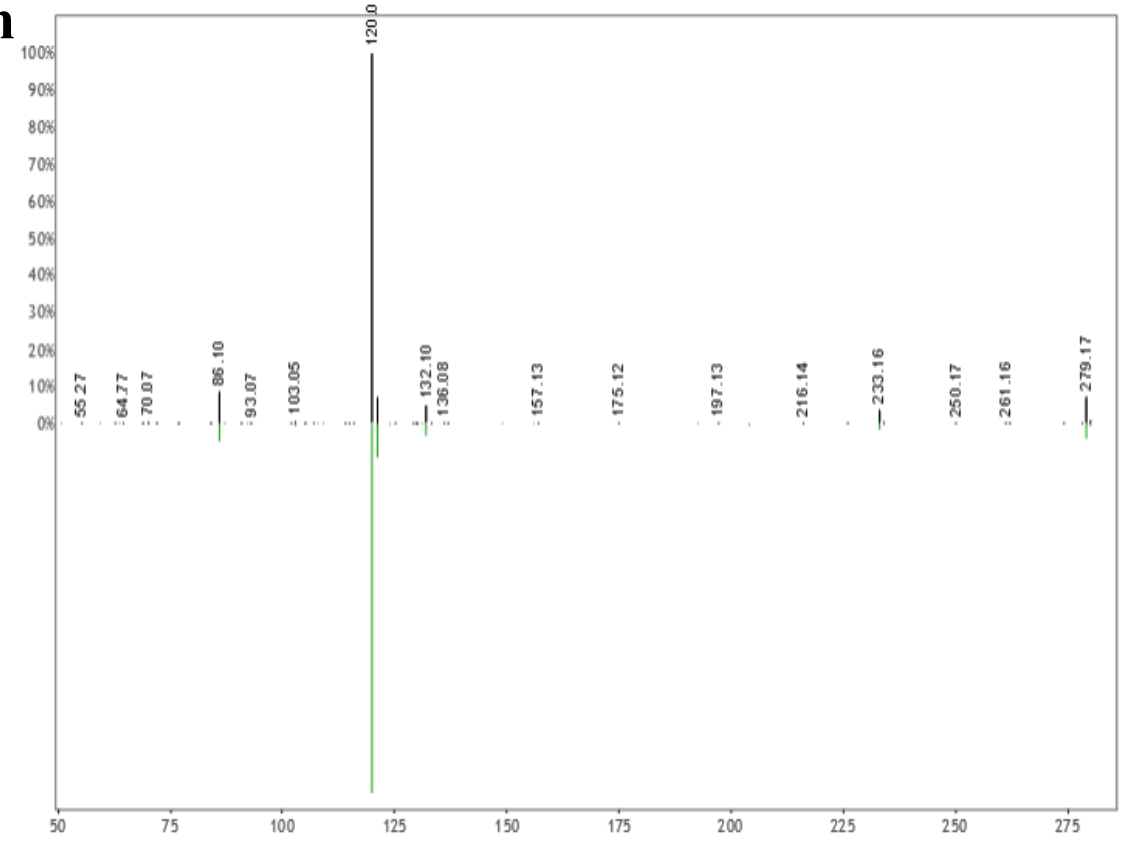
af


272.1
70.07
60.06
84.08
95.06
110.07
116.07
137.05
140.08
157.11
166.03
175.12
194.13
198.99
213.12
226.12
238.12
245.00
255.15
289.02
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
75
100
125
150
175
200
225
250
275

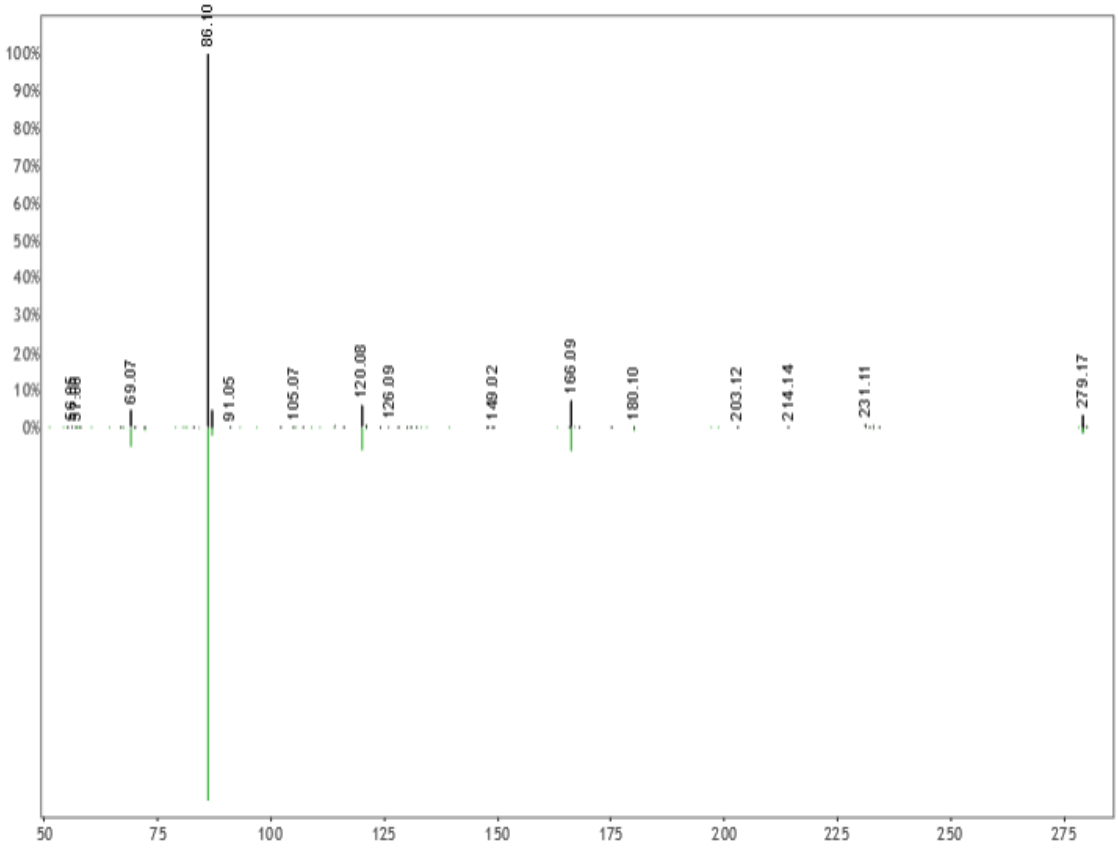
ag


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
67.05
81.07
91.05
105.07
119.09
131.09
133.10
147.12
159.12
173.13
183.12
197.13
201.13
229.19
239.18
247.21
257.19
275.2
50
75
100
125
150
175
200
225
250
275

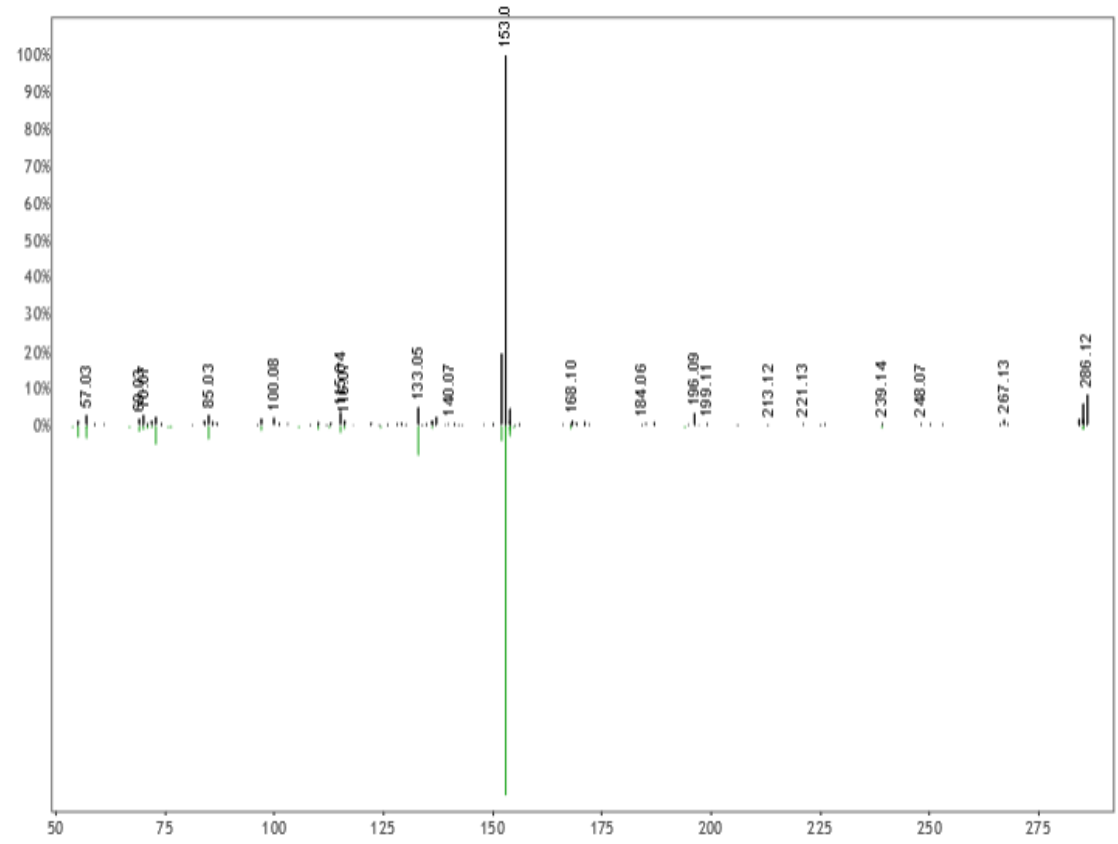
ah


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.27
64.77
70.07
86.10
93.07
103.05
120.0
132.10
136.08
157.13
175.12
197.13
216.14
233.16
250.17
261.16
279.17
50
75
100
125
150
175
200
225
250
275

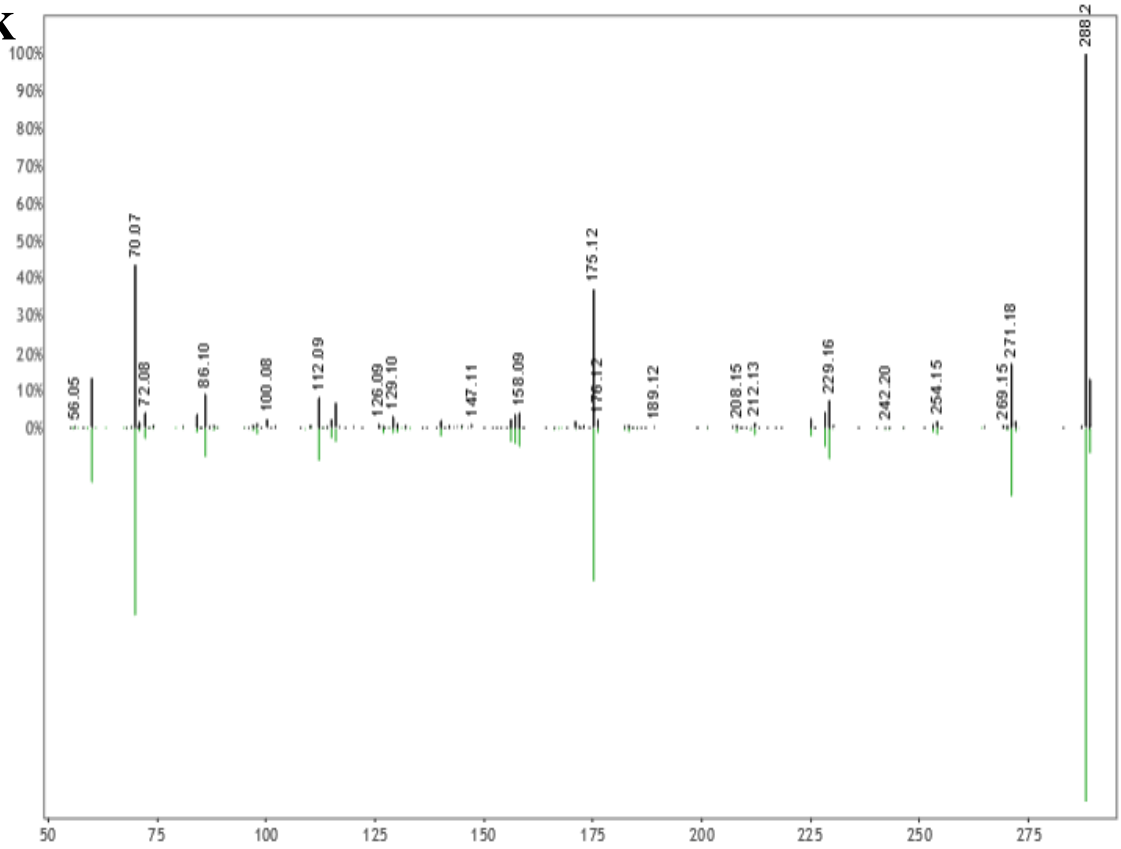
ai


86.10
69.07
91.05
105.07
120.08
126.09
149.02
166.09
180.10
203.12
214.14
231.11
279.17
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
75
100
125
150
175
200
225
250
275

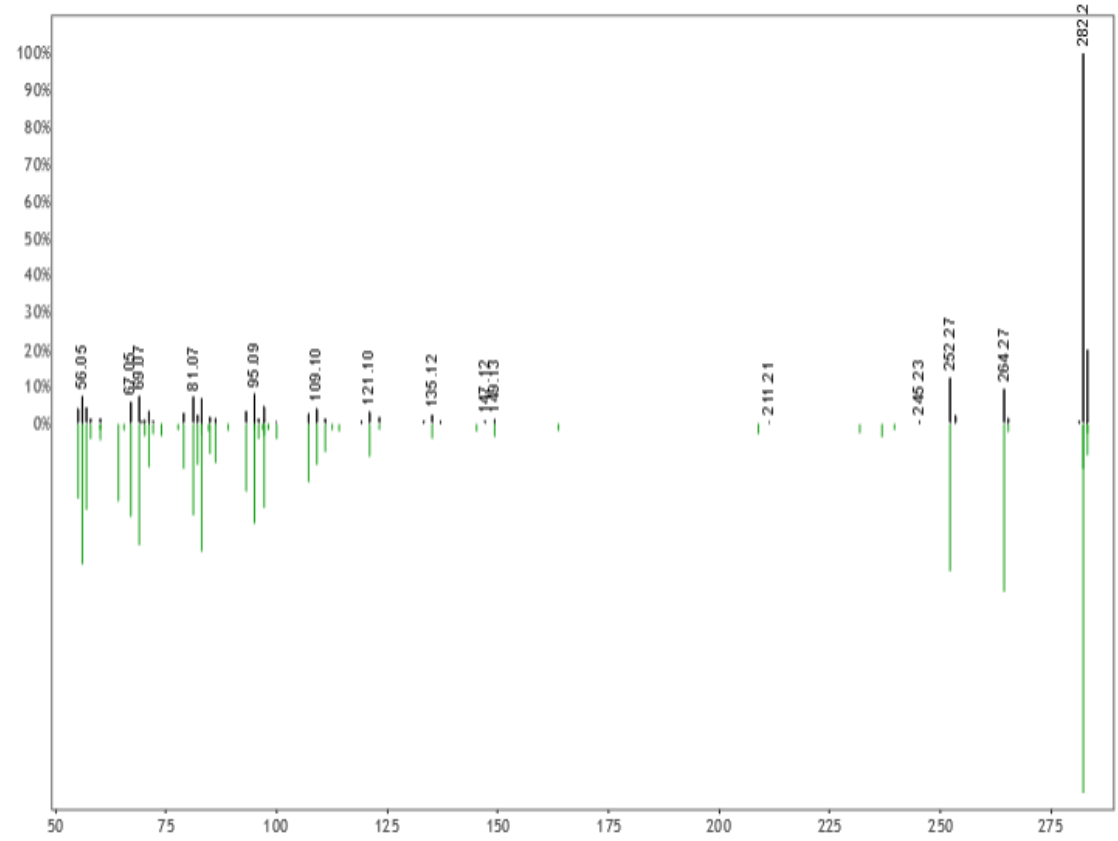
aj


153.0
57.03
85.03
100.08
133.05
140.07
168.10
184.06
196.09
199.11
213.12
221.13
239.14
248.07
267.13
286.12
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
75
100
125
150
175
200
225
250
275

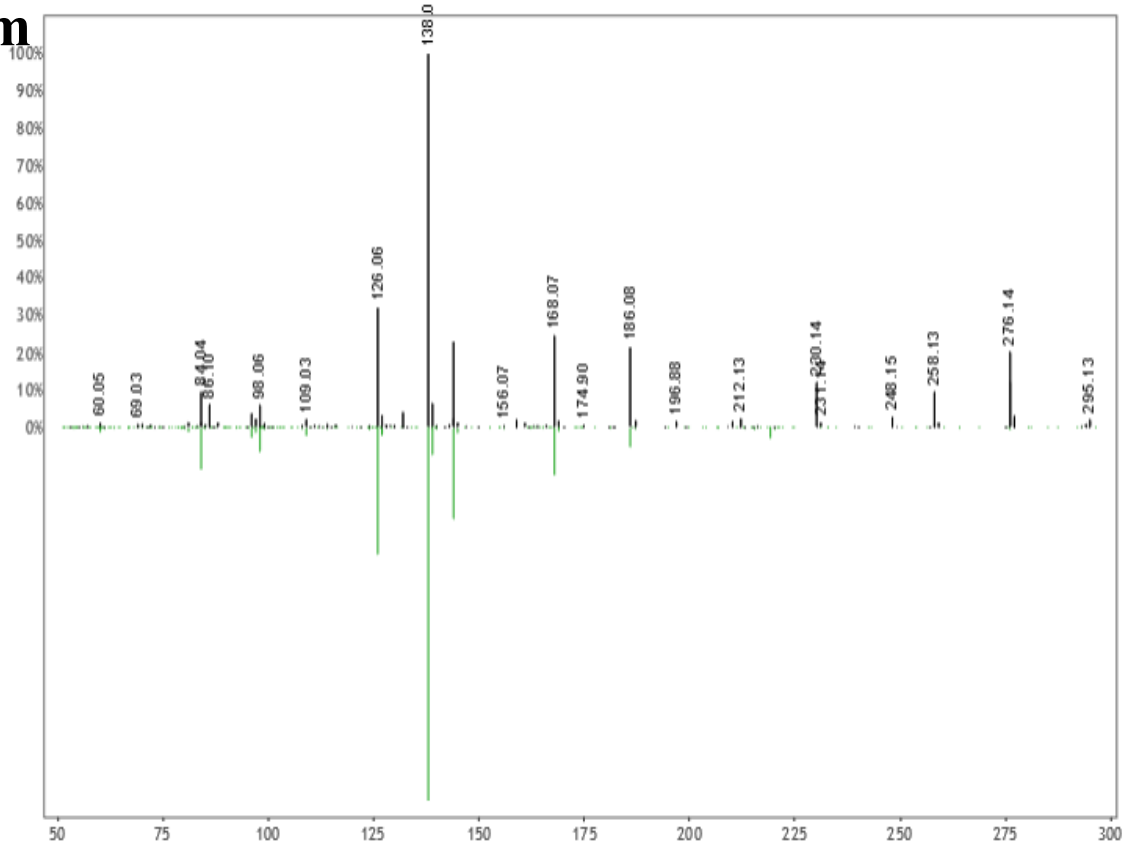
ak


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
70.07
72.08
86.10
100.08
112.09
126.09
129.10
147.11
158.09
175.12
176.12
189.12
208.15
212.13
229.16
242.20
254.15
269.15
271.18
288.2
50
75
100
125
150
175
200
225
250
275

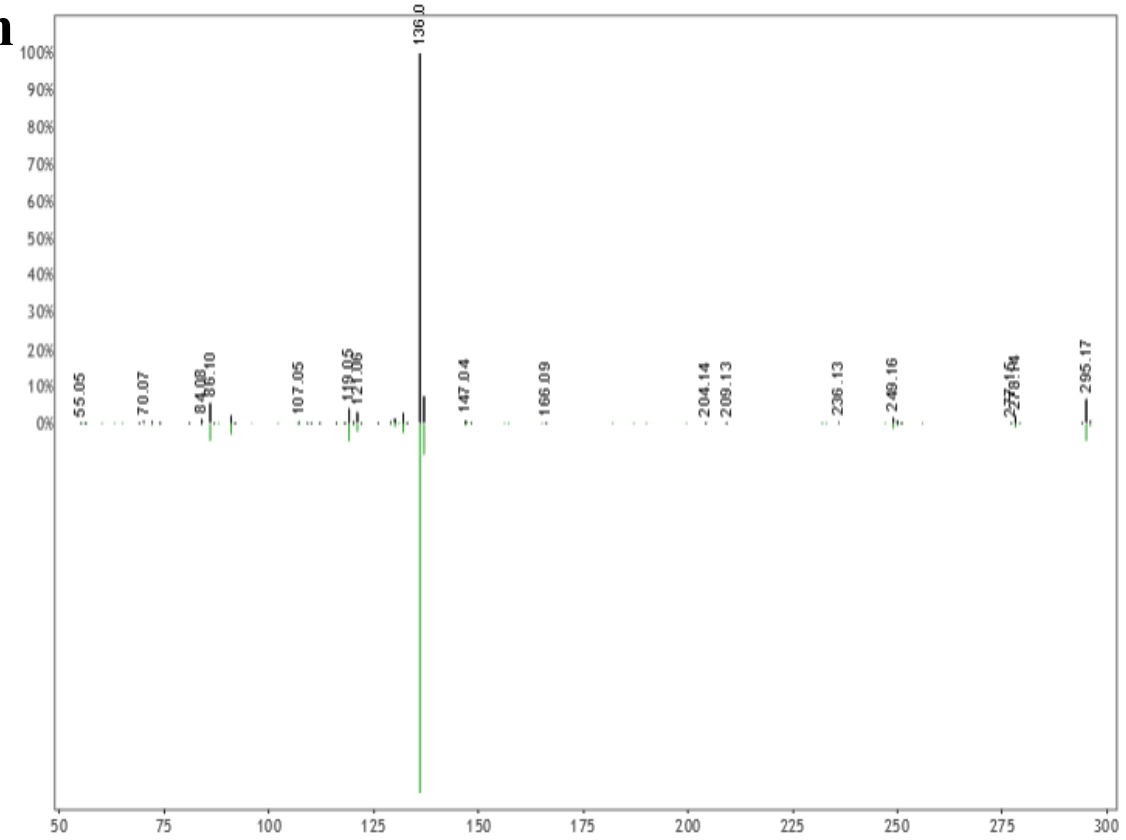
al


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
67.05
69.07
81.07
95.09
109.10
121.10
135.12
211.21
245.23
252.27
264.27
282.2
50
75
100
125
150
175
200
225
250
275

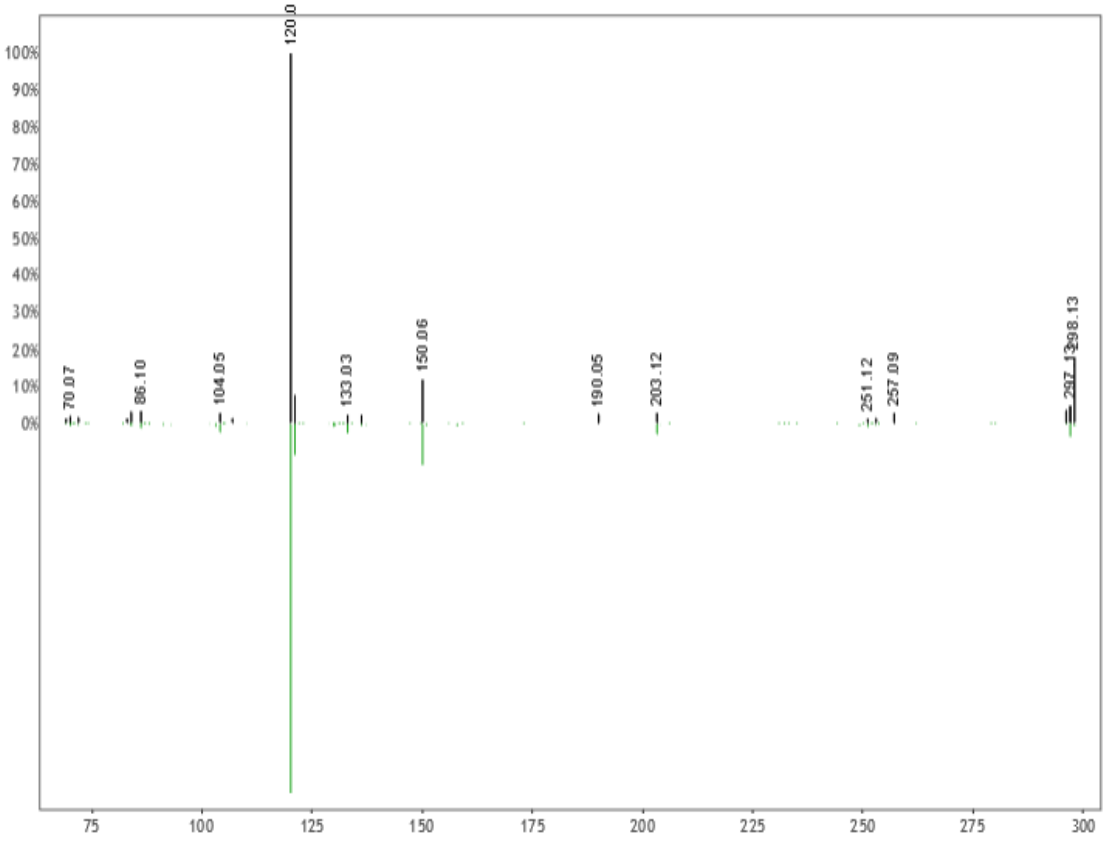
am


138.0
126.06
168.07
186.08
276.14
60.05
69.03
84.04
86.10
98.06
109.03
156.07
174.90
196.88
212.13
248.15
258.13
295.13
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
75
100
125
150
175
200
225
250
275
300

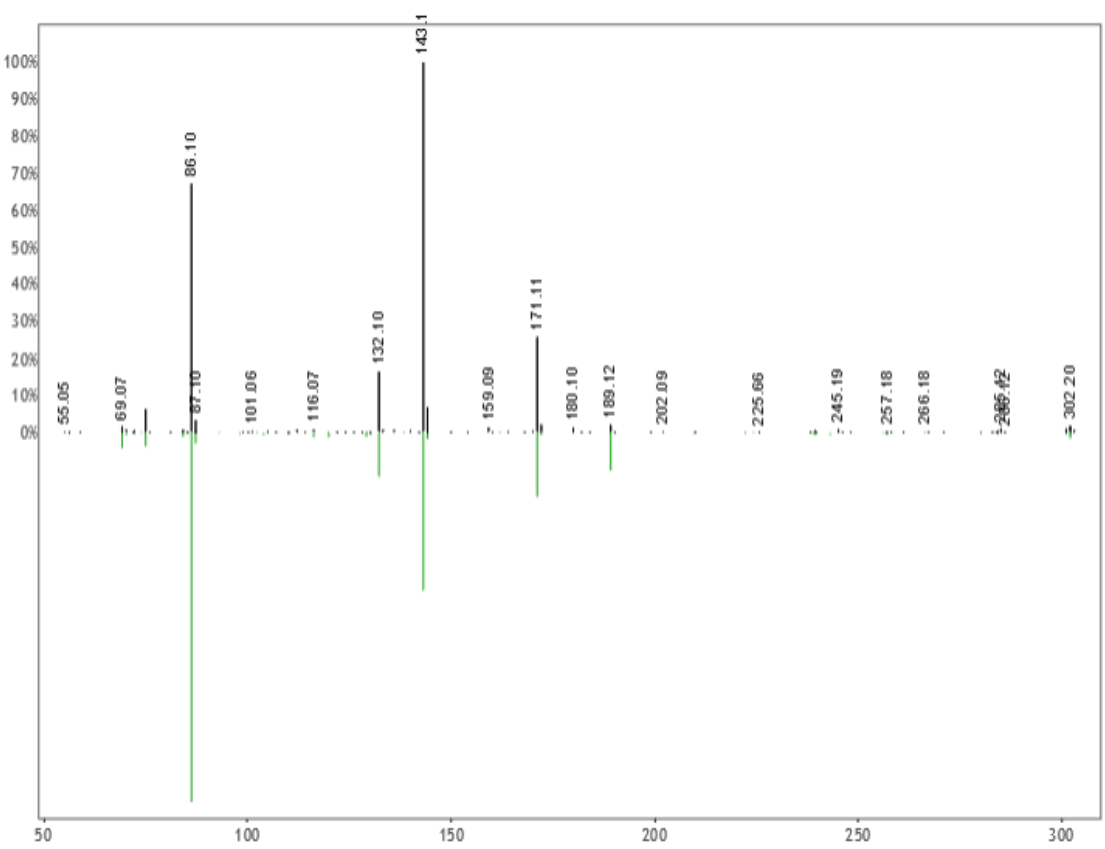
an


136.0
55.05
70.07
84.08
86.10
107.05
119.05
121.08
147.04
166.09
204.14
209.13
236.13
249.16
295.17
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
75
100
125
150
175
200
225
250
275
300

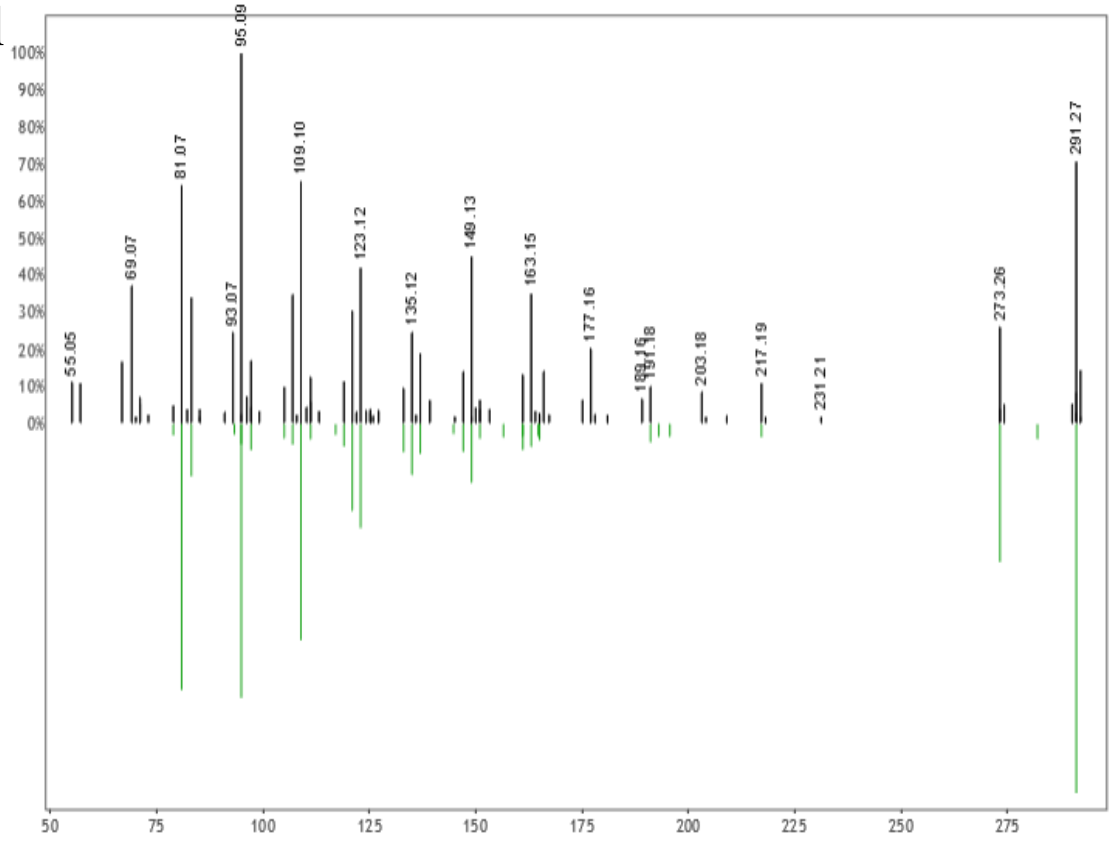
ao


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
70.07
86.10
104.05
120.0
133.03
150.06
190.05
203.12
251.12
257.09
298.13
75
100
125
150
175
200
225
250
275
300

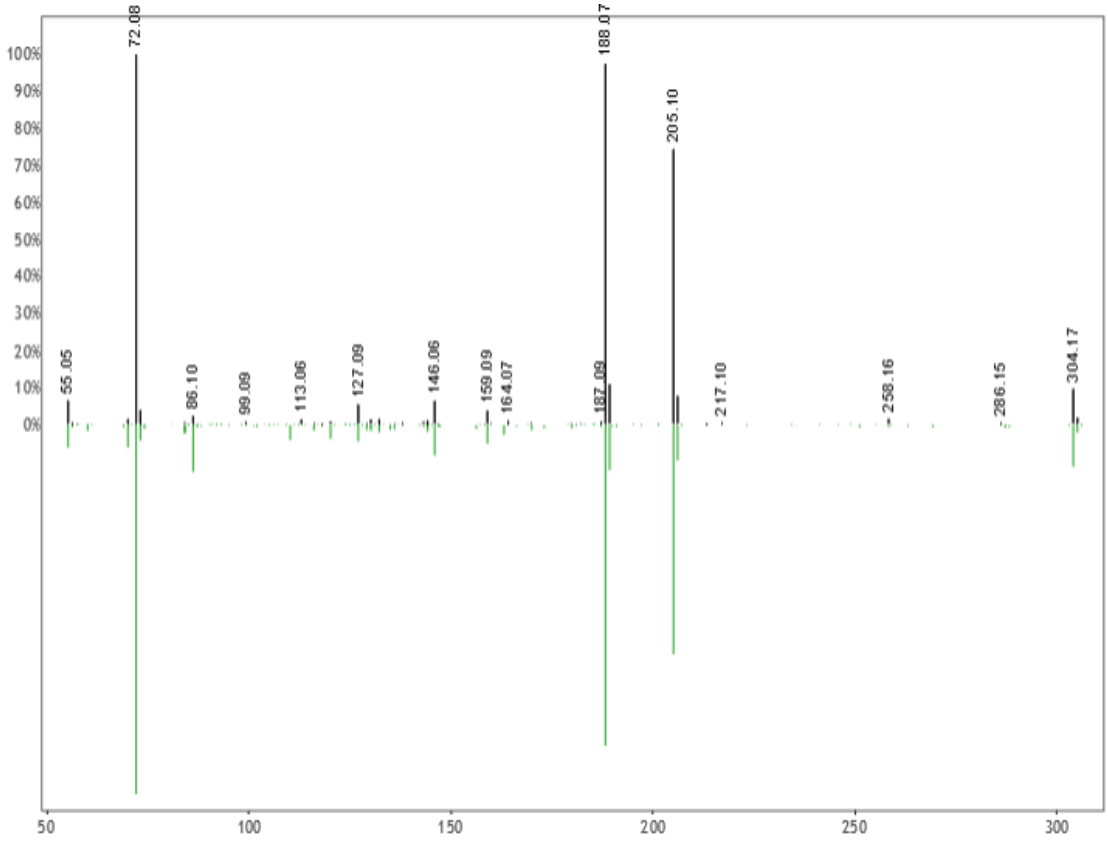
ap


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
69.07
86.10
87.10
101.06
116.07
132.10
143.1
159.09
171.11
180.10
189.12
202.09
225.66
245.19
257.18
266.18
302.20
50
100
150
200
250
300

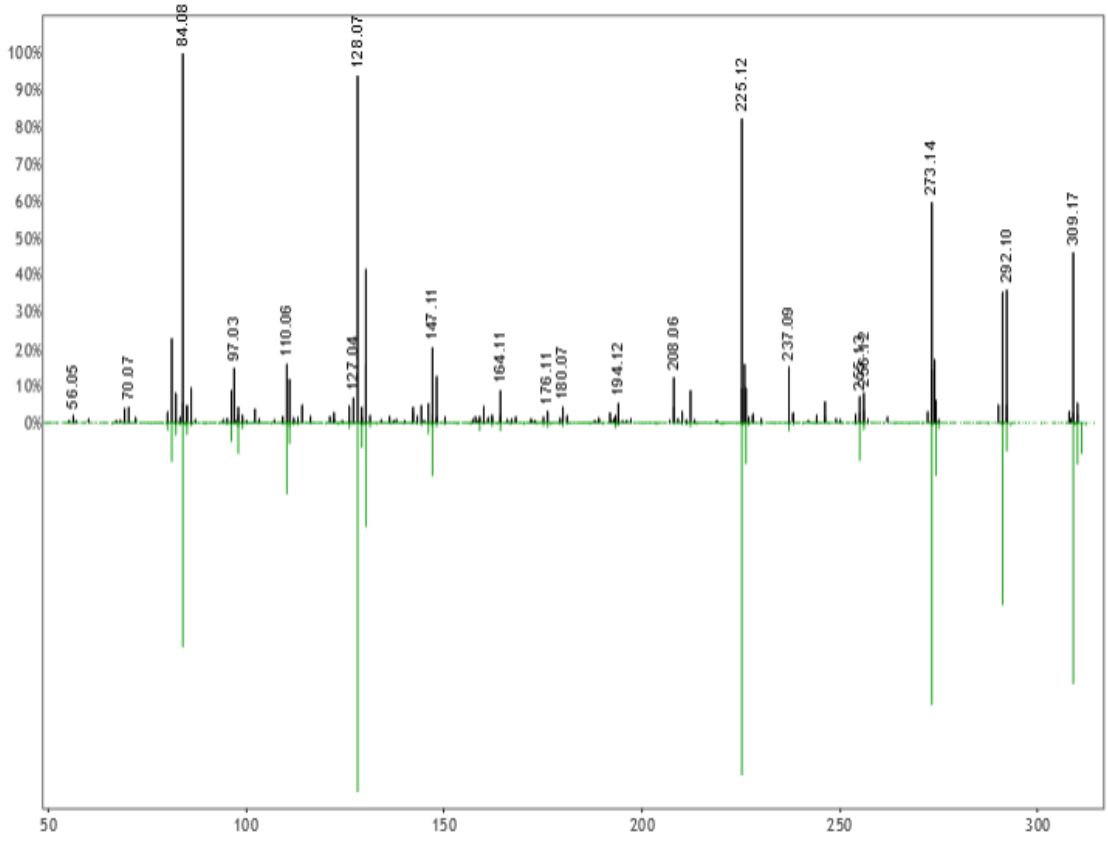
aq


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
69.07
81.07
93.07
95.09
109.10
123.12
135.12
149.13
163.15
177.16
189.16
191.18
203.18
217.19
231.21
273.26
291.27
50
75
100
125
150
175
200
225
250
275

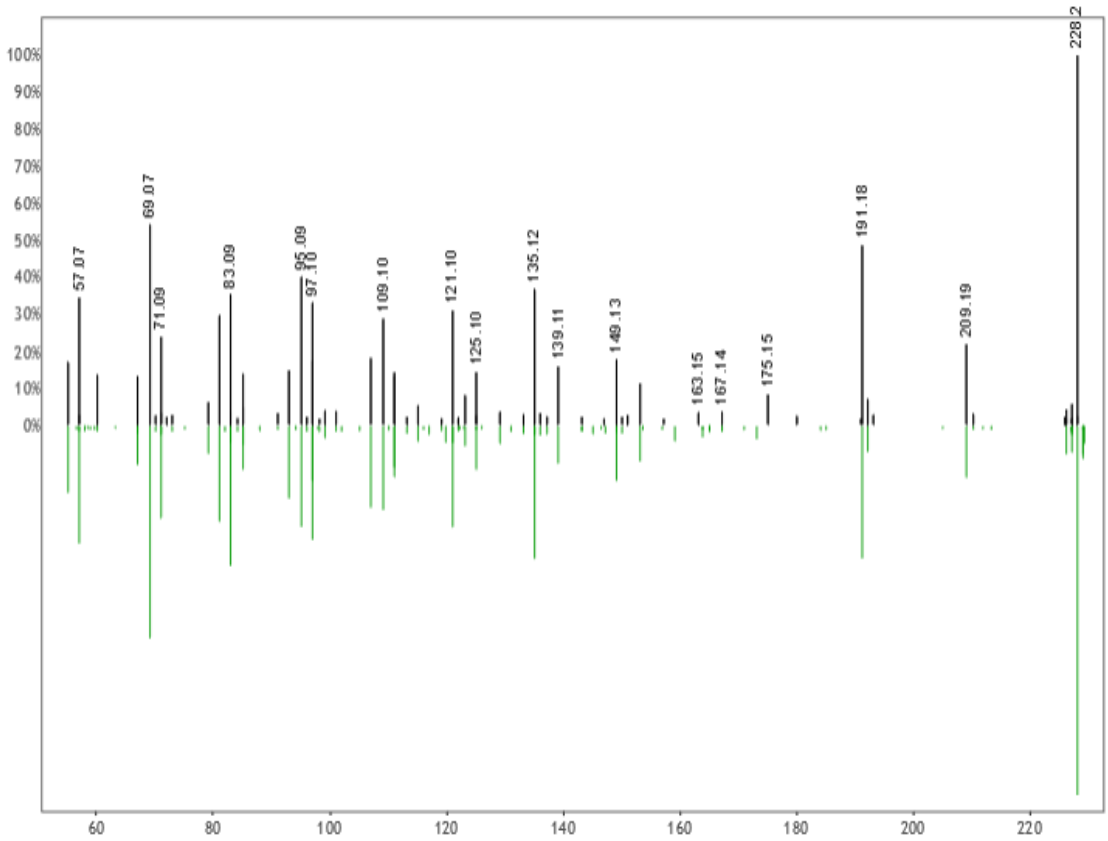
ar


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
72.08
86.10
99.09
113.06
127.09
146.06
159.09
164.07
187.09
188.07
205.10
217.10
258.16
286.15
304.17
50
100
150
200
250
300

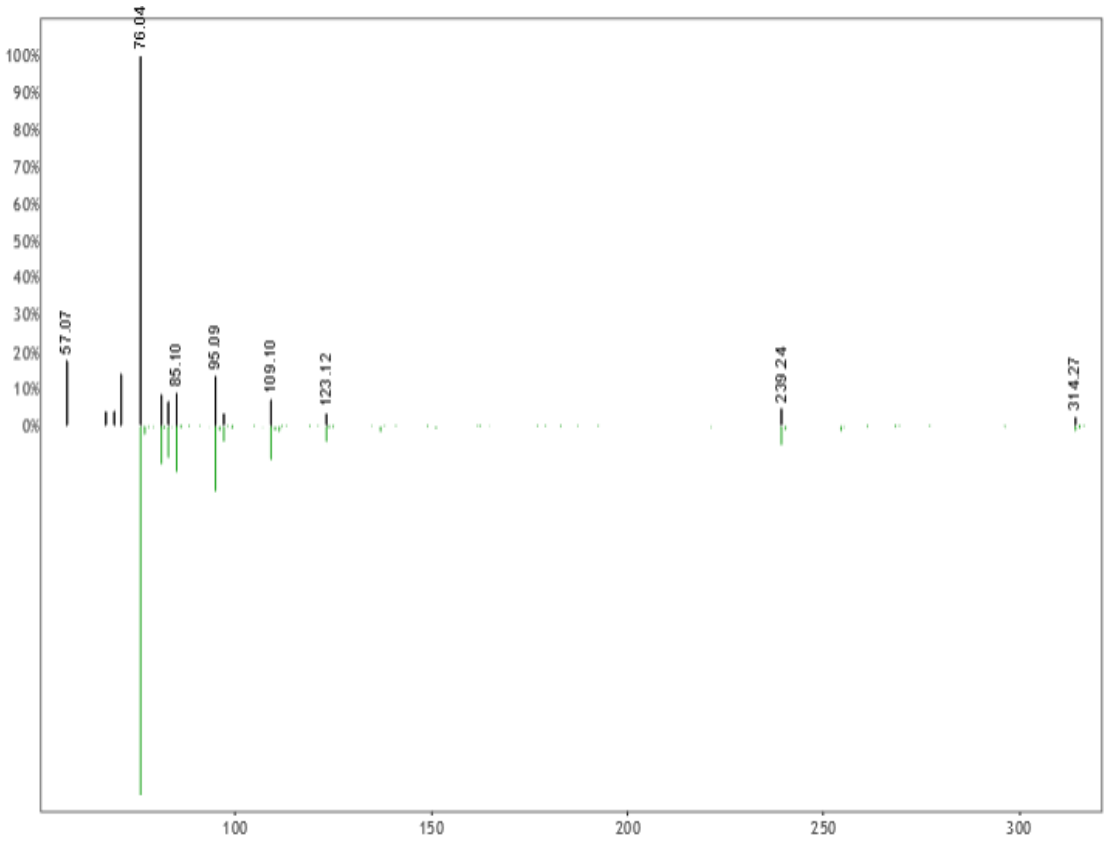
as


84.08
128.07
225.12
273.14
309.17
292.10
147.11
110.06
97.03
56.05
70.07
127.04
164.11
176.11
180.07
194.12
208.06
237.09
50
100
150
200
250
300

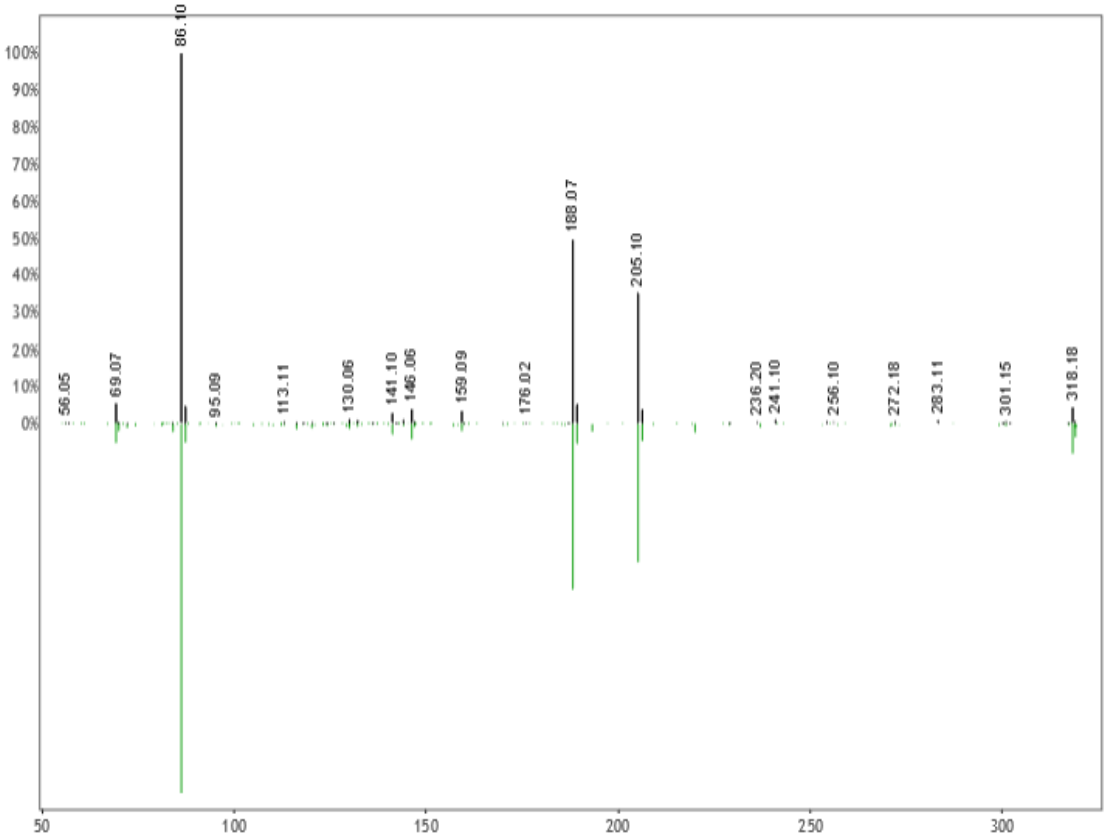
at


228.2
69.07
57.07
71.09
83.09
95.09
97.10
109.10
121.10
125.10
135.12
139.11
149.13
163.15
167.14
175.15
191.18
209.19
60
80
100
120
140
160
180
200
220

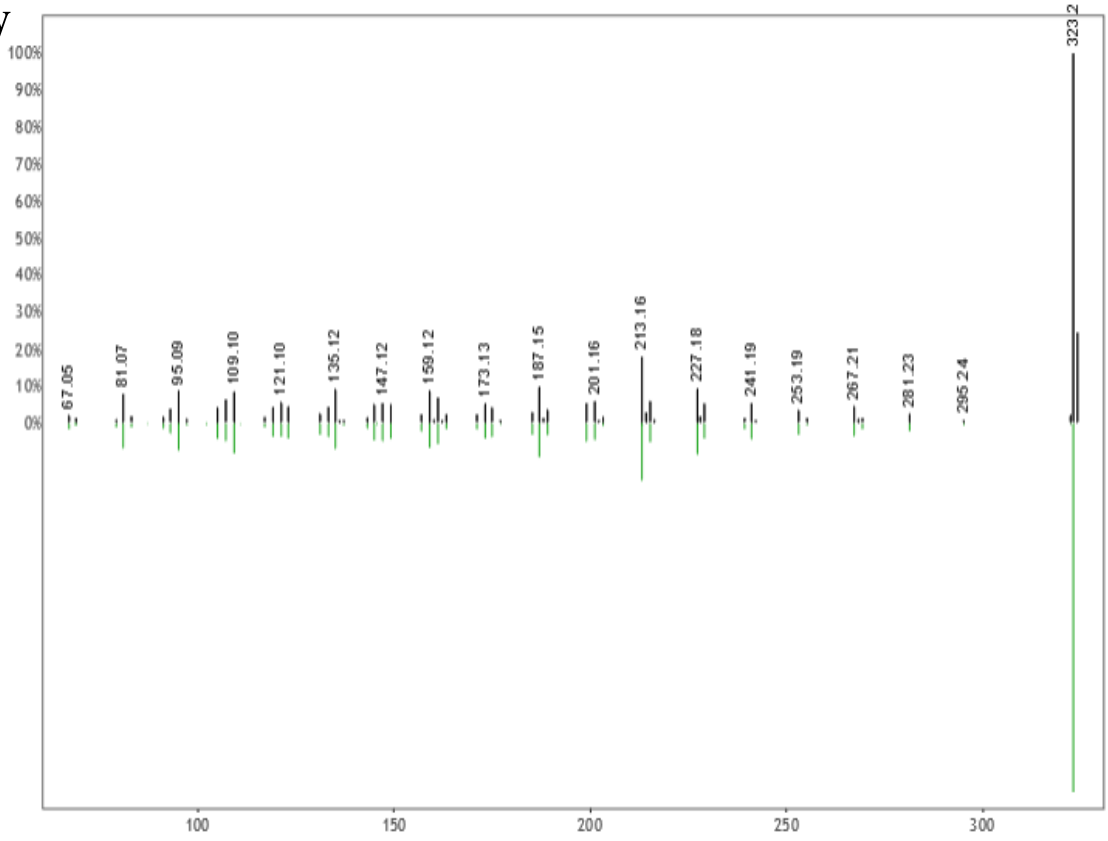
au


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
76.04
85.10
95.09
109.10
123.12
239.24
314.27
100
150
200
250
300

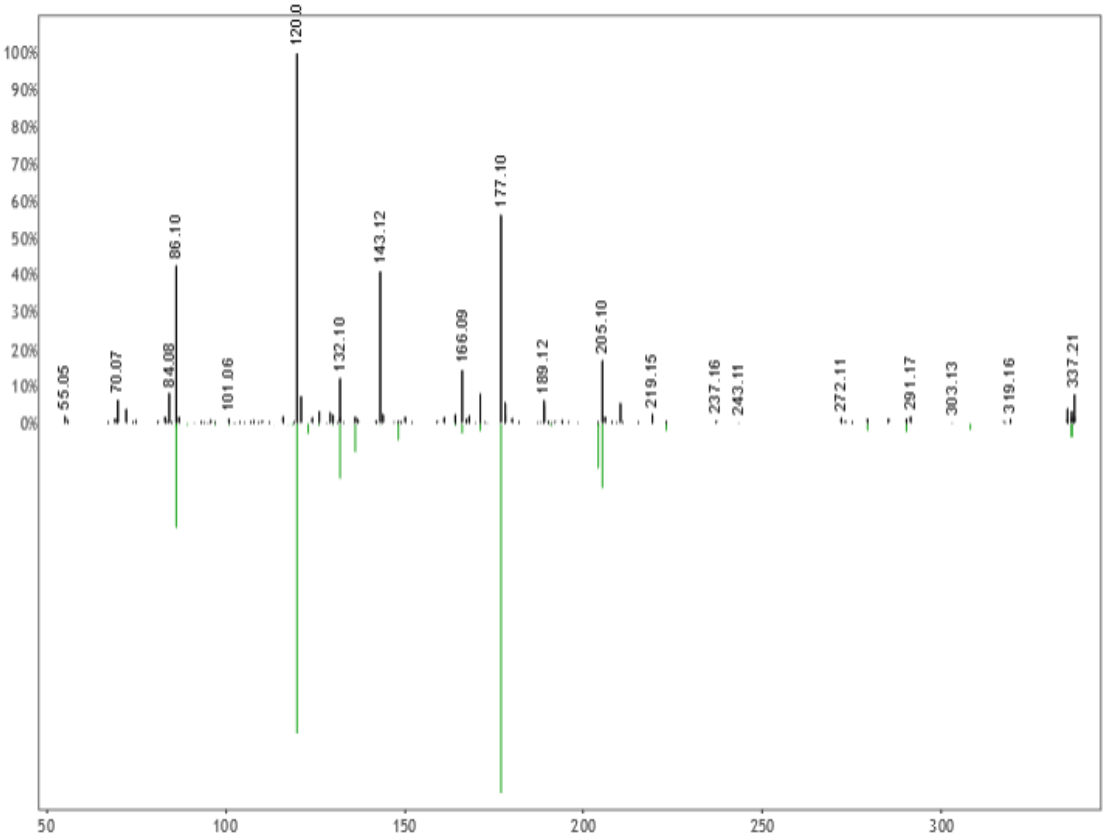
av


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56.05
69.07
86.10
95.09
113.11
130.06
141.10
146.06
159.09
176.02
188.07
205.10
236.20
241.10
256.10
272.18
283.11
301.15
318.18
50
100
150
200
250
300

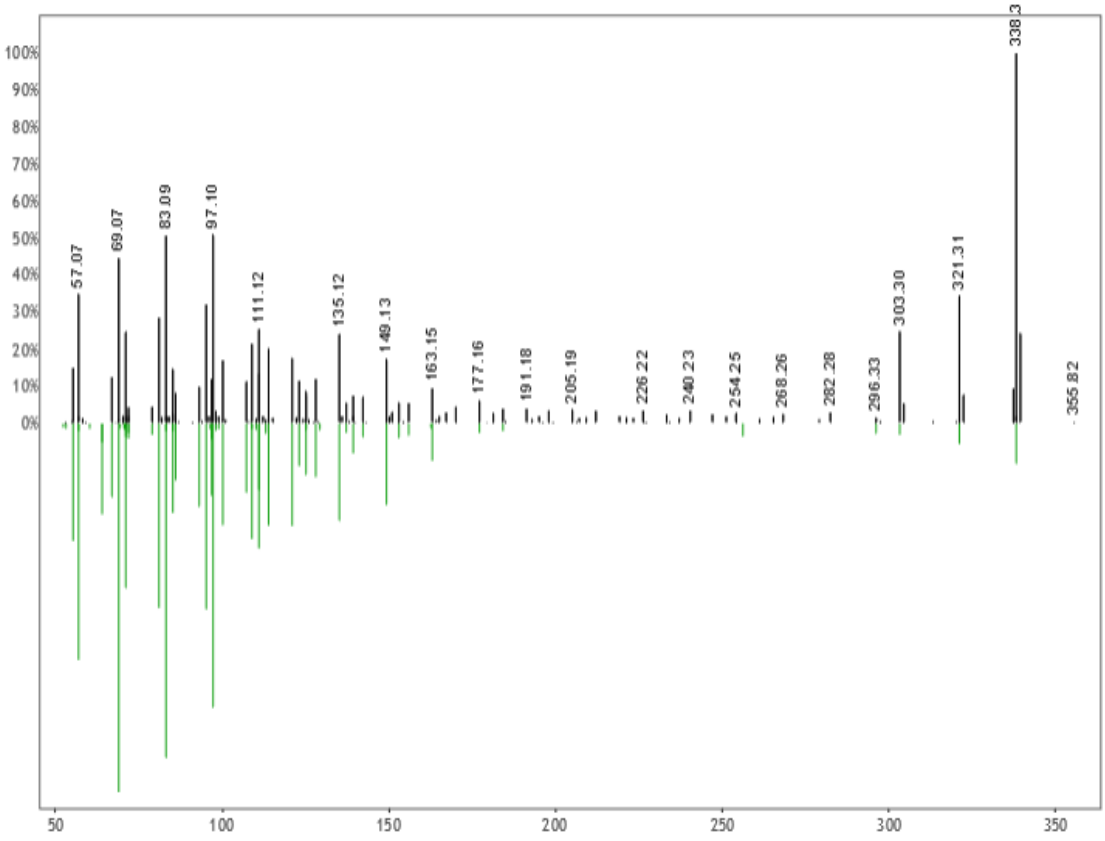
aw


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
67.05
81.07
95.09
109.10
121.10
135.12
147.12
159.12
173.13
187.15
201.16
213.16
227.18
241.19
253.19
267.21
281.23
295.24
323.2
100
150
200
250
300

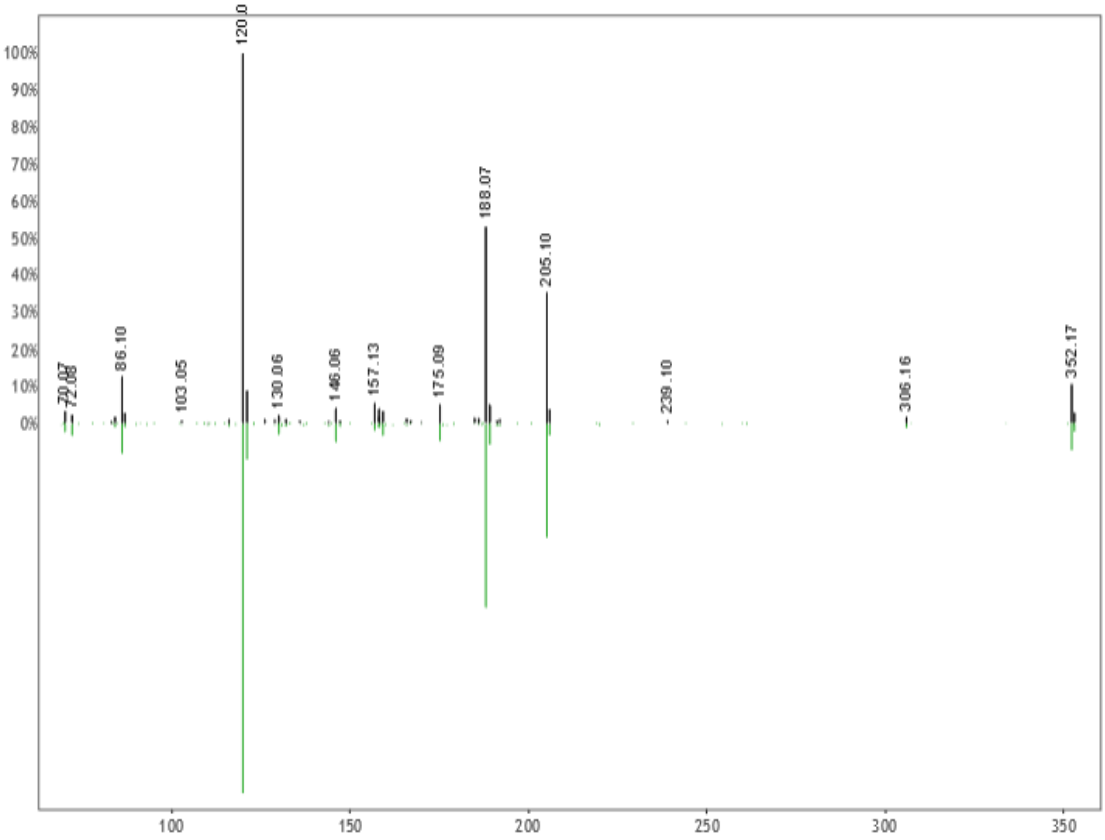
ax


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
70.07
84.08
86.10
101.06
120.0
132.10
143.12
166.09
177.10
189.12
205.10
219.15
237.16
243.11
272.11
291.17
303.13
319.16
337.21
50
100
150
200
250
300

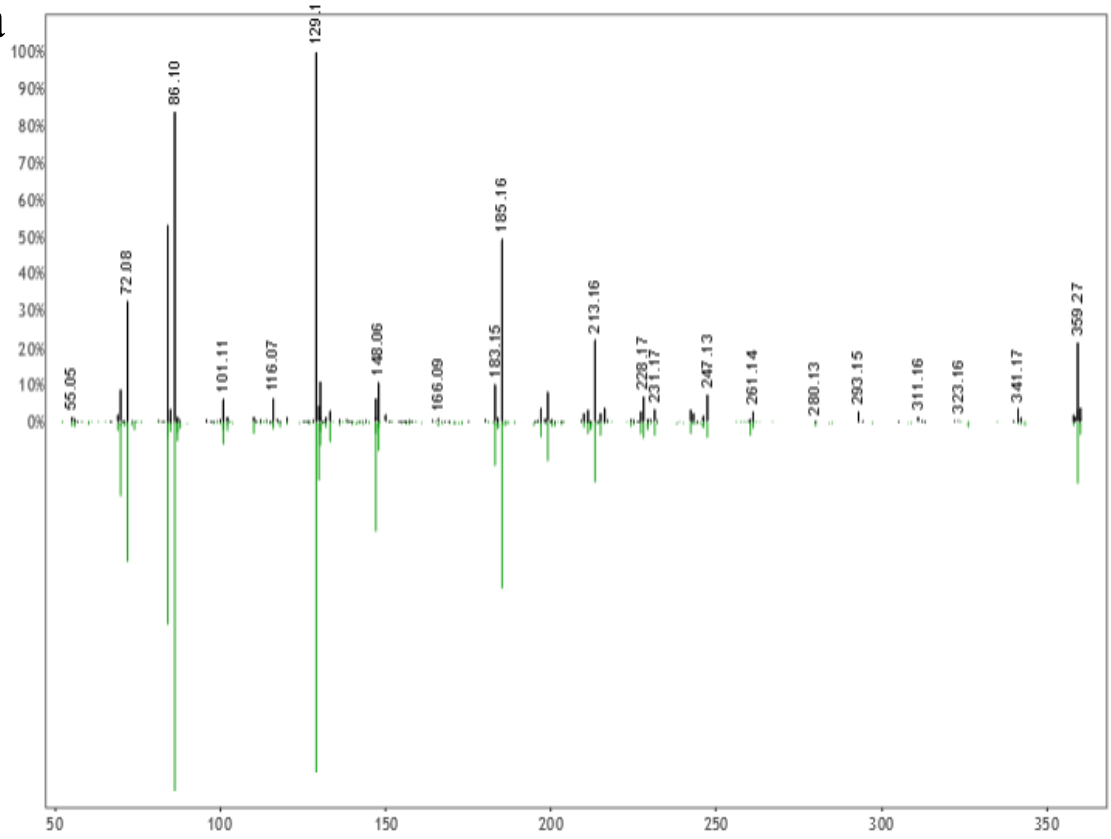
ay


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
69.07
83.09
97.10
111.12
135.12
149.13
163.15
177.16
191.18
205.19
226.22
240.23
254.25
268.26
282.28
296.33
303.30
321.31
338.3
355.82
50
100
150
200
250
300
350

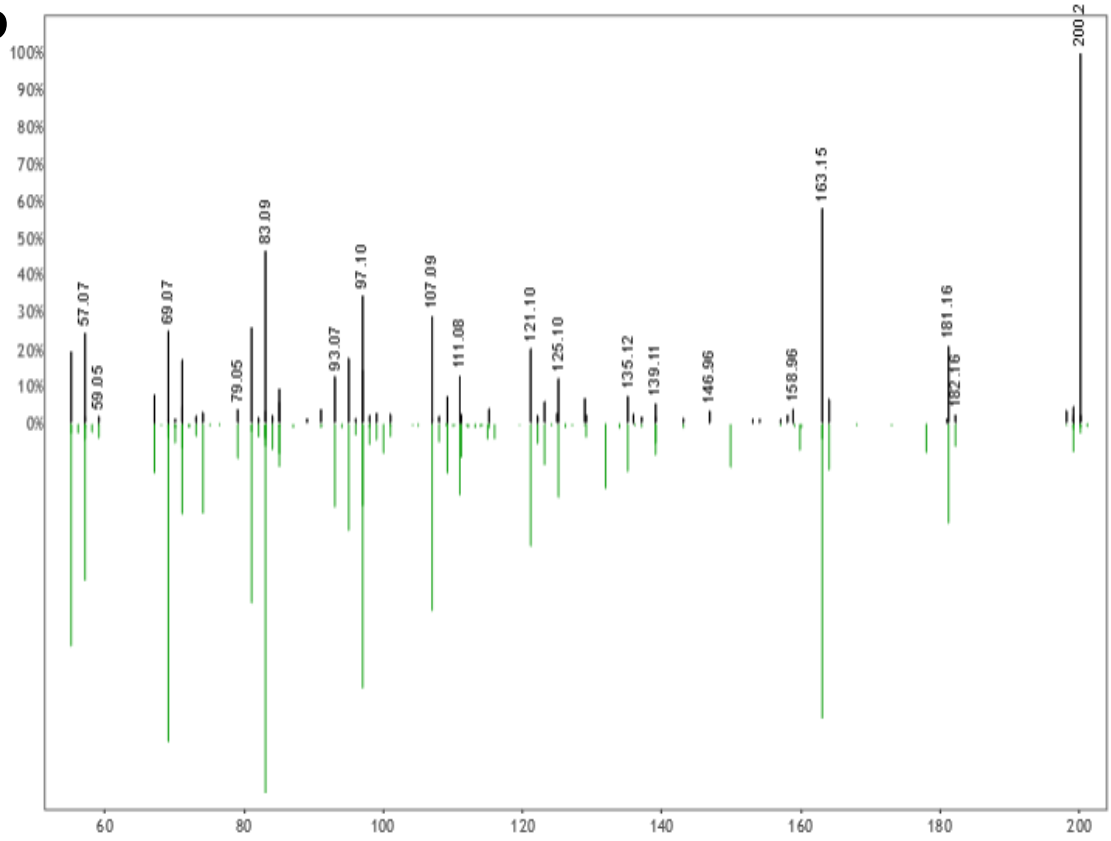
az


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
72.08
86.10
103.05
120.0
130.06
146.06
157.13
175.09
188.07
205.10
239.10
306.16
352.17
100
150
200
250
300
350

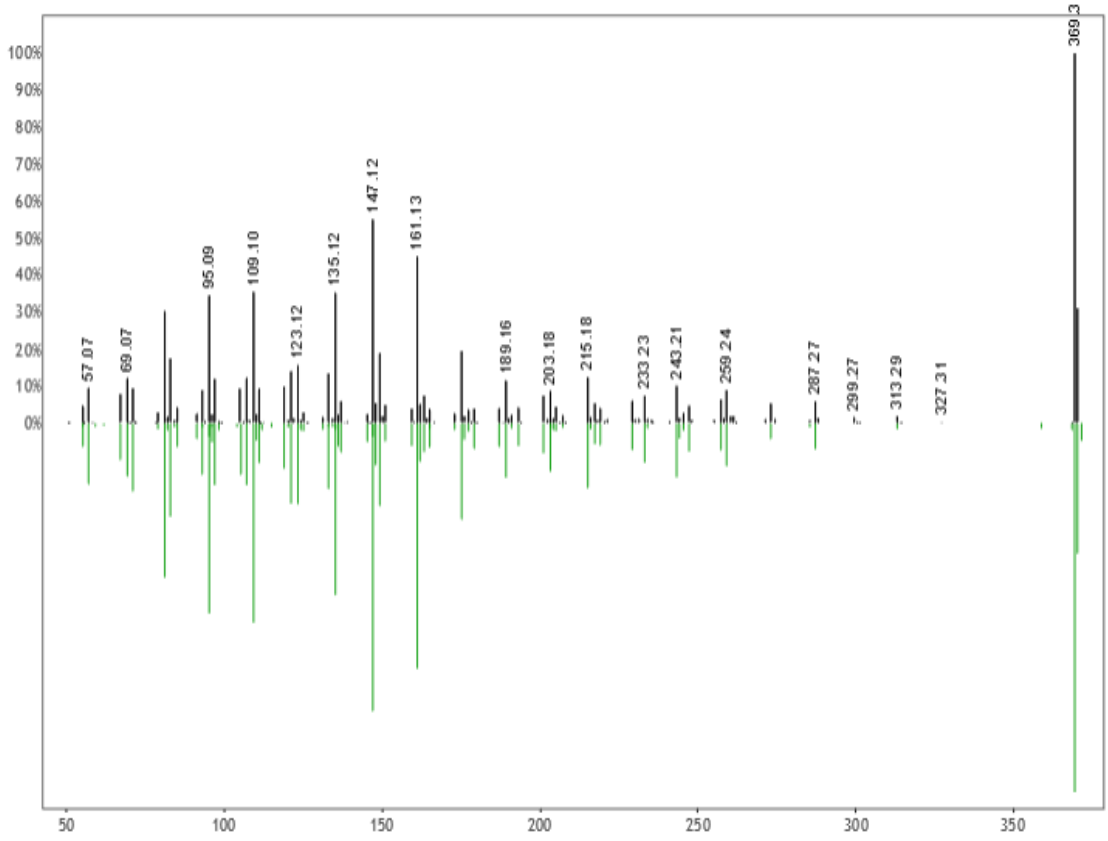
ba


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
72.08
86.10
101.11
116.07
129.1
148.06
166.09
183.15
185.16
213.16
226.17
231.17
247.13
261.14
280.13
293.15
311.16
323.16
341.17
359.27
50
100
150
200
250
300
350

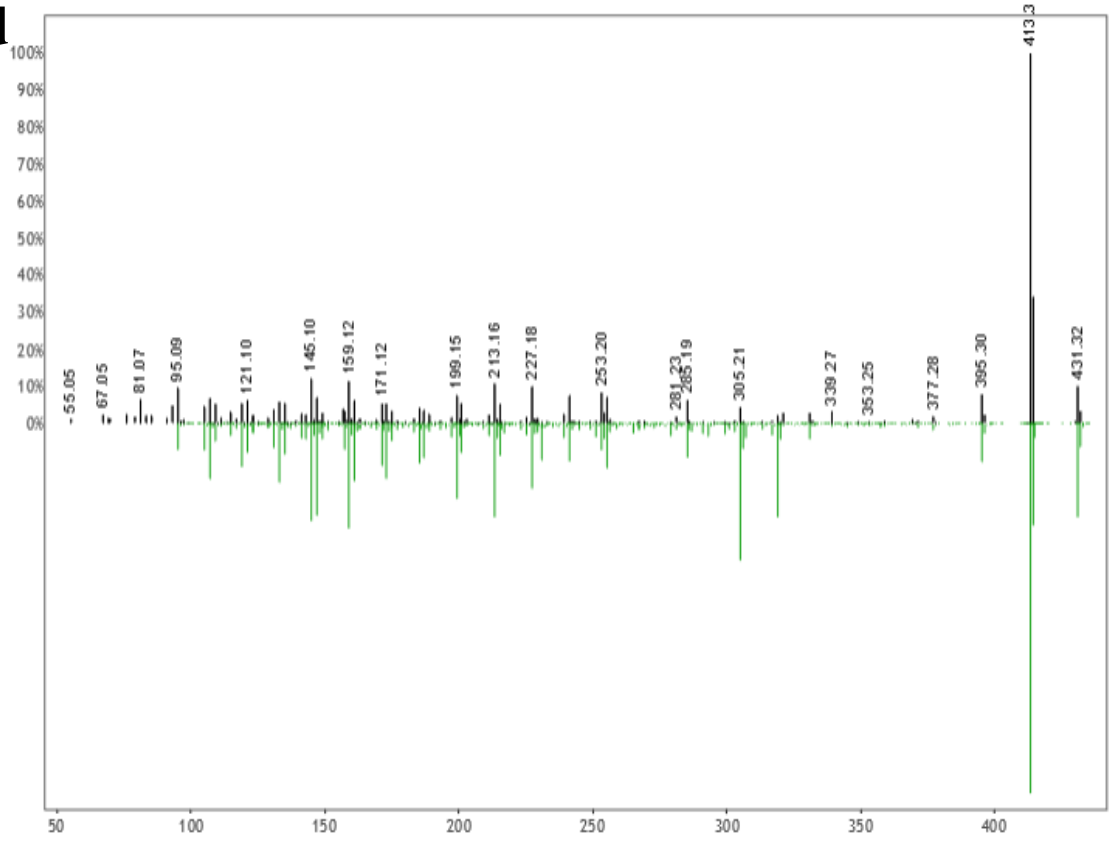
bb


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
59.05
69.07
79.05
83.09
93.07
97.10
107.09
111.08
121.10
125.10
135.12
139.11
146.96
158.96
163.15
181.16
182.16
200.2
60
80
100
120
140
160
180
200

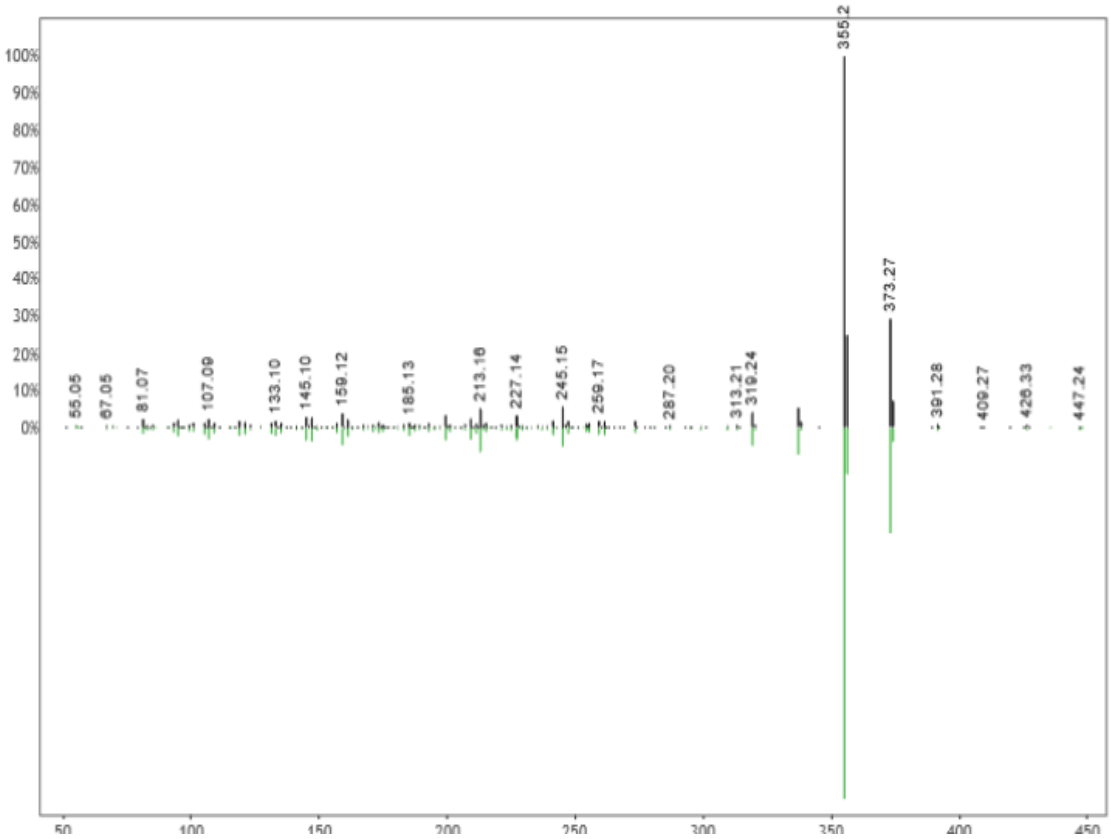
bc


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
69.07
95.09
109.10
123.12
135.12
147.12
161.13
189.16
203.18
215.18
233.23
243.21
259.24
287.27
299.27
313.29
327.31
369.3
50
100
150
200
250
300
350

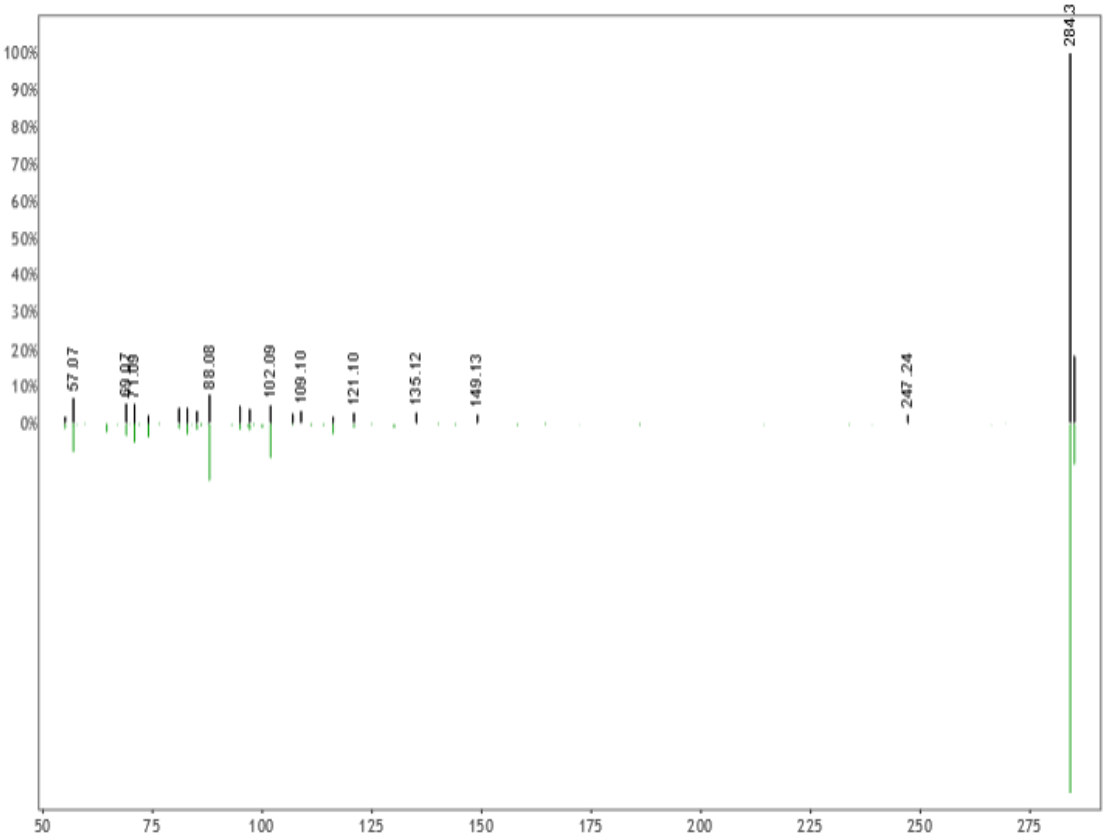
bd


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
67.05
81.07
95.09
121.10
145.10
159.12
171.12
199.15
213.16
227.18
253.20
281.23
285.19
305.21
339.27
353.25
377.28
395.30
413.3
431.32
50
100
150
200
250
300
350
400

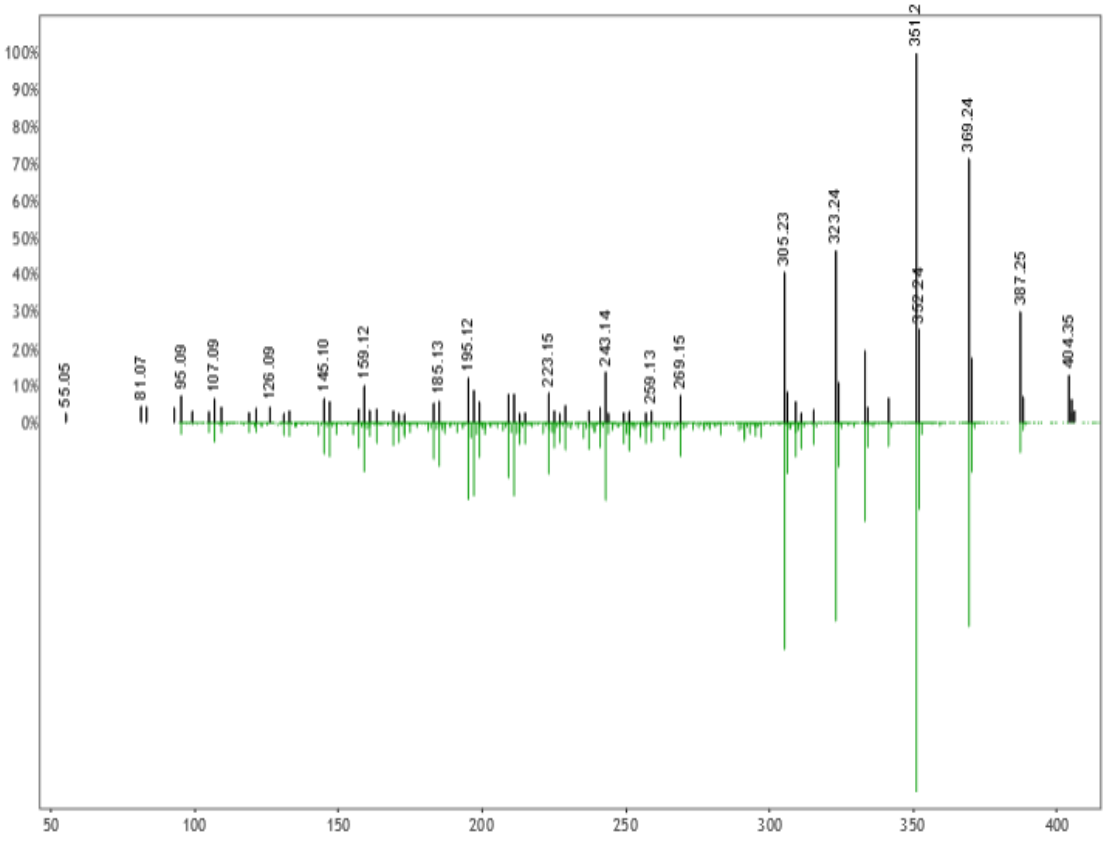
be


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
67.05
81.07
107.09
133.10
145.10
159.12
185.13
213.16
227.14
245.15
259.17
287.20
313.21
319.24
355.2
373.27
391.28
409.27
426.33
447.24
50
100
150
200
250
300
350
400
450

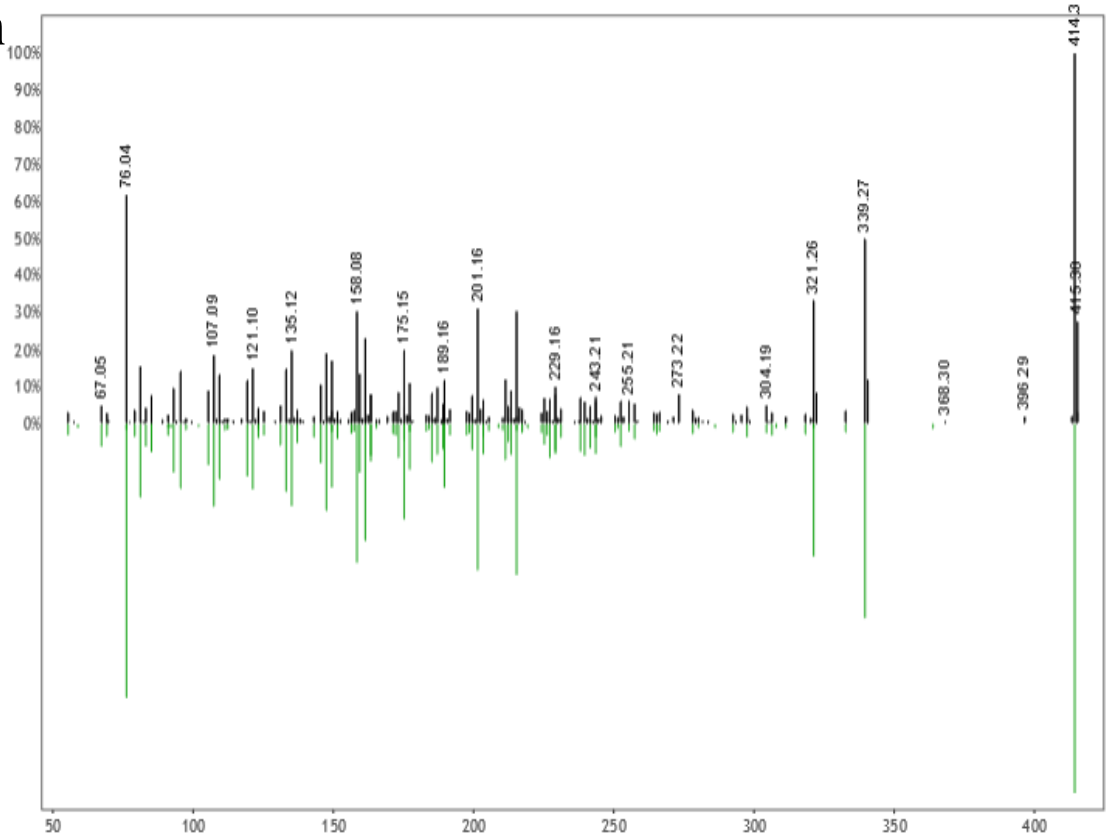
bf


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
57.07
88.08
102.09
109.10
121.10
135.12
149.13
247.24
284.3
50
75
100
125
150
175
200
225
250
275

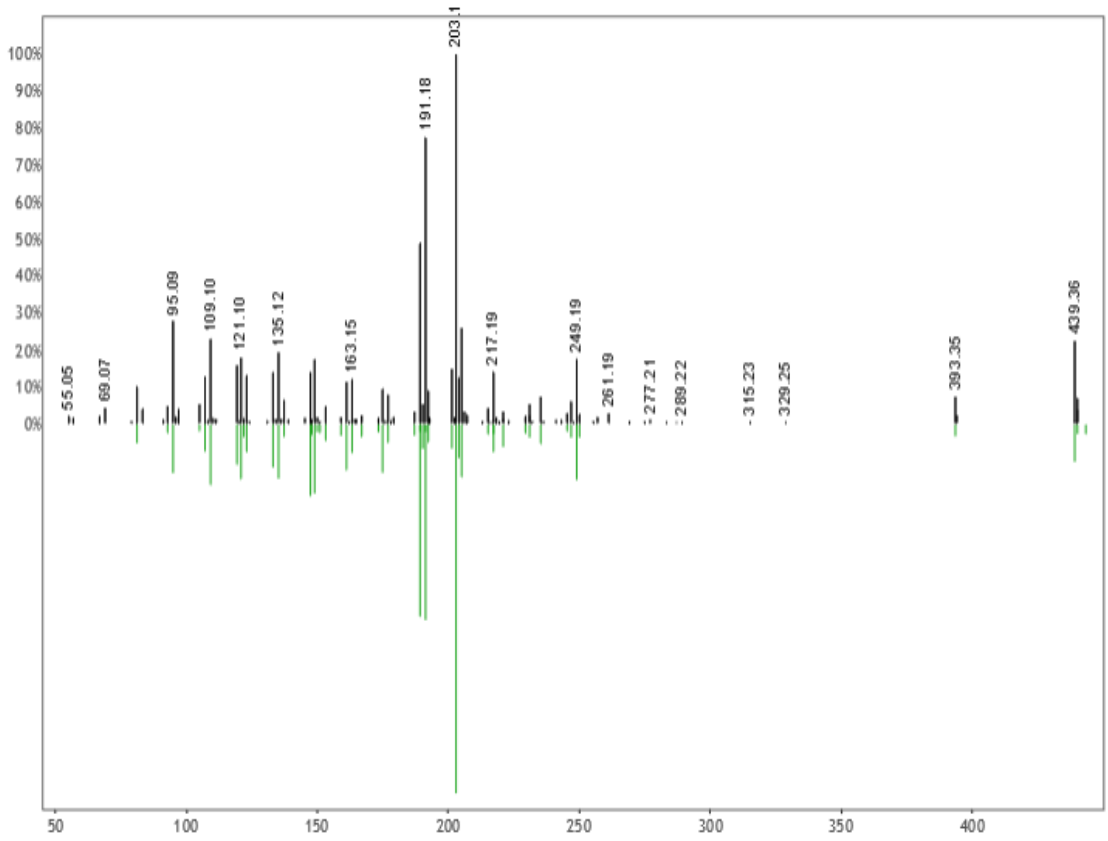
bg


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
81.07
95.09
107.09
126.09
145.10
159.12
185.13
195.12
223.15
243.14
259.13
269.15
305.23
323.24
351.2
352.24
369.24
387.25
404.35
50
100
150
200
250
300
350
400

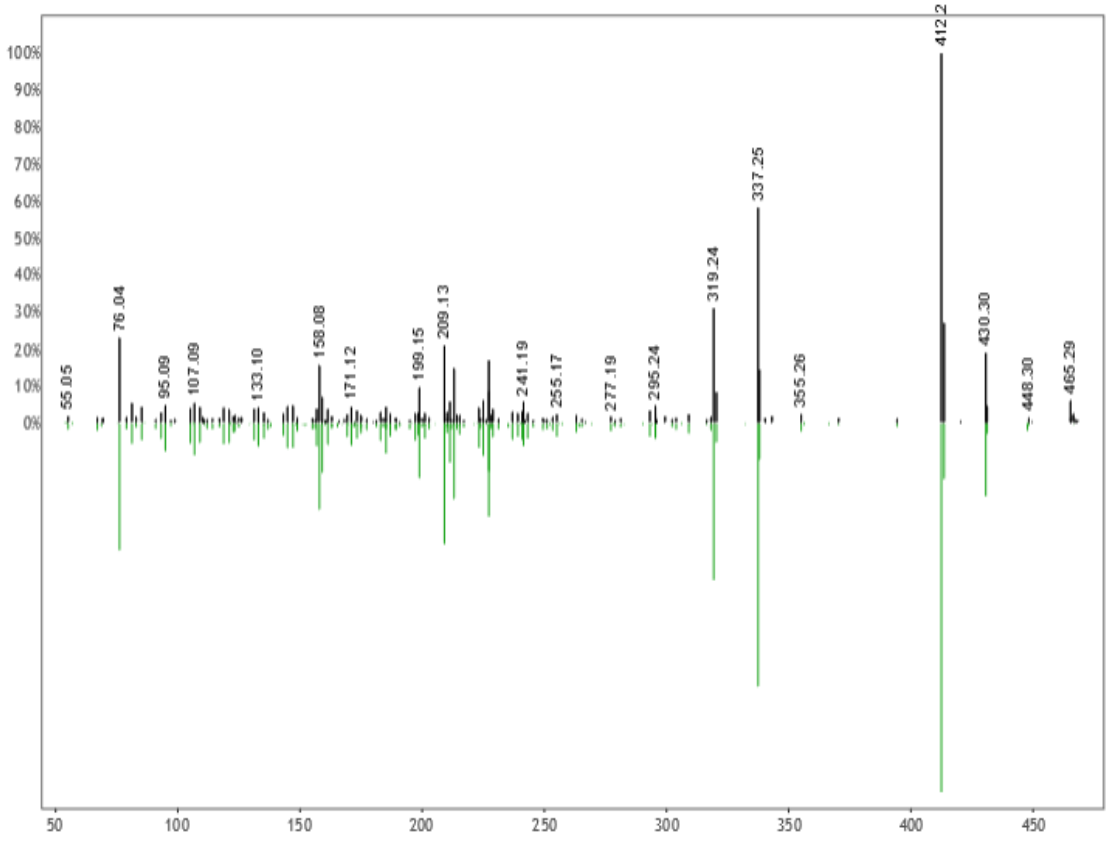
bh


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
67.05
76.04
107.09
121.10
135.12
158.08
175.15
189.16
201.16
229.16
243.21
255.21
273.22
304.19
321.26
339.27
368.30
396.29
414.3
415.30
50
100
150
200
250
300
350
400

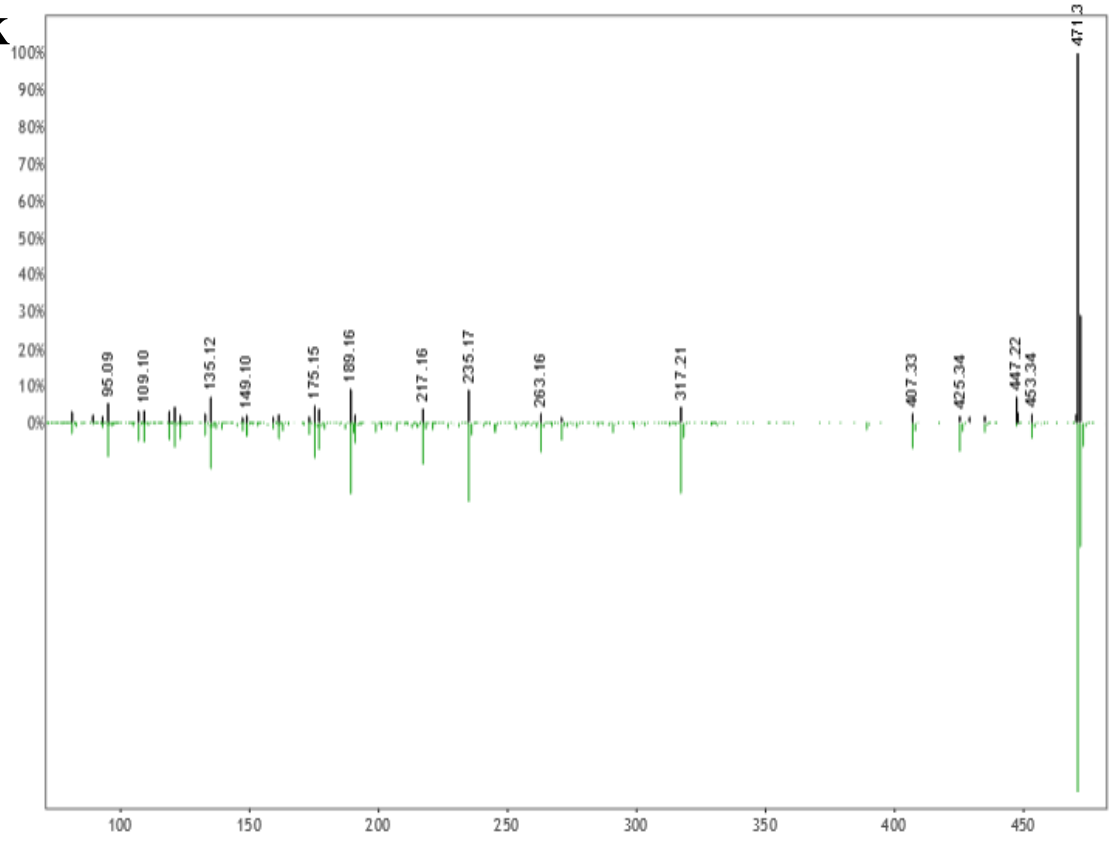
bi


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
69.07
95.09
109.10
121.10
135.12
163.15
191.18
203.1
217.19
249.19
261.19
277.21
289.22
315.23
329.25
393.35
439.36
50
100
150
200
250
300
350
400

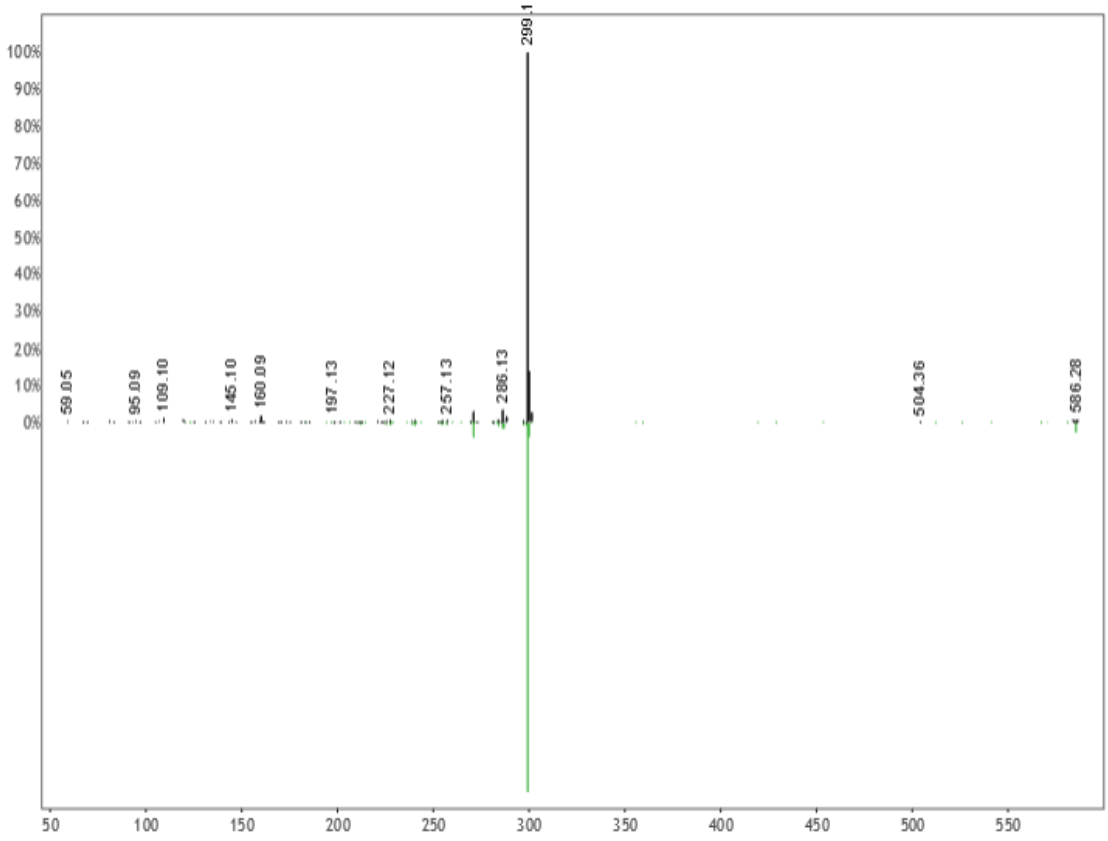
bj


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
55.05
76.04
95.09
107.09
133.10
158.08
171.12
199.15
209.13
241.19
255.17
277.19
295.24
319.24
337.25
355.26
412.2
430.30
448.30
465.29
50
100
150
200
250
300
350
400
450

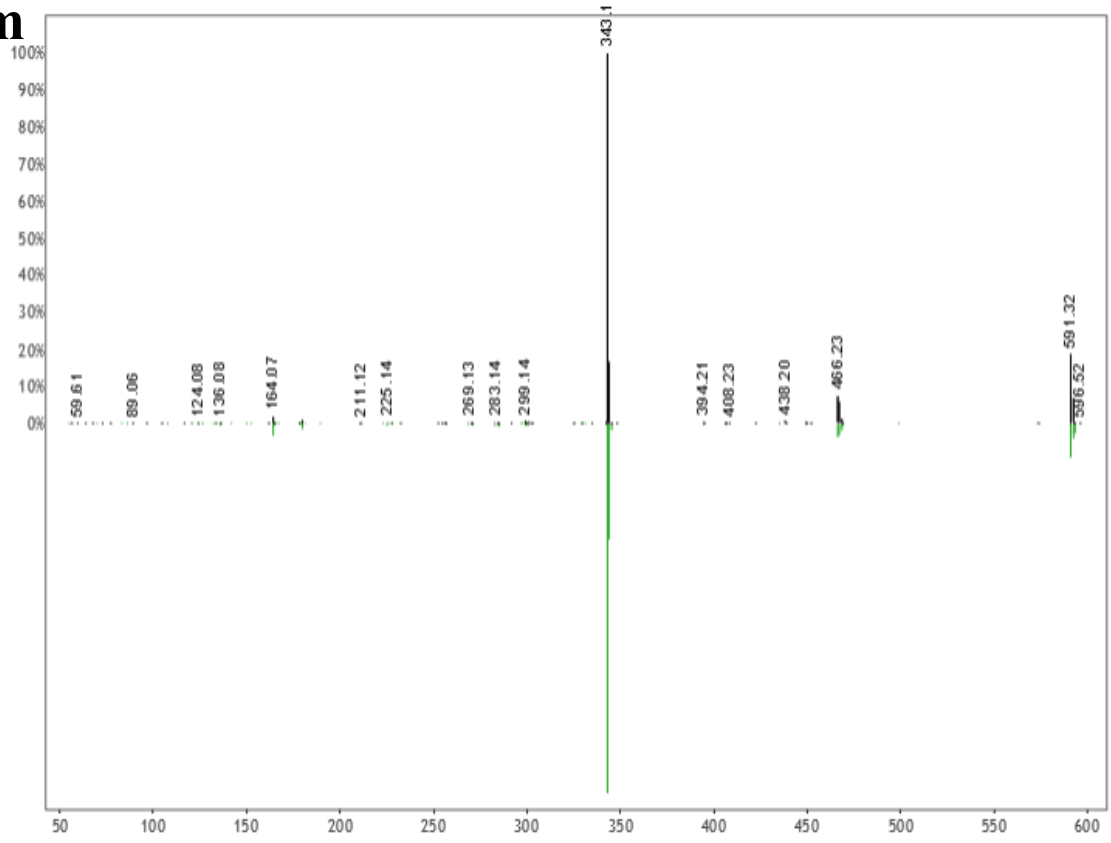
bk


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
95.09
109.10
135.12
149.10
175.15
189.16
217.16
235.17
263.16
317.21
407.33
425.34
447.22
453.34
471.3
100
150
200
250
300
350
400
450

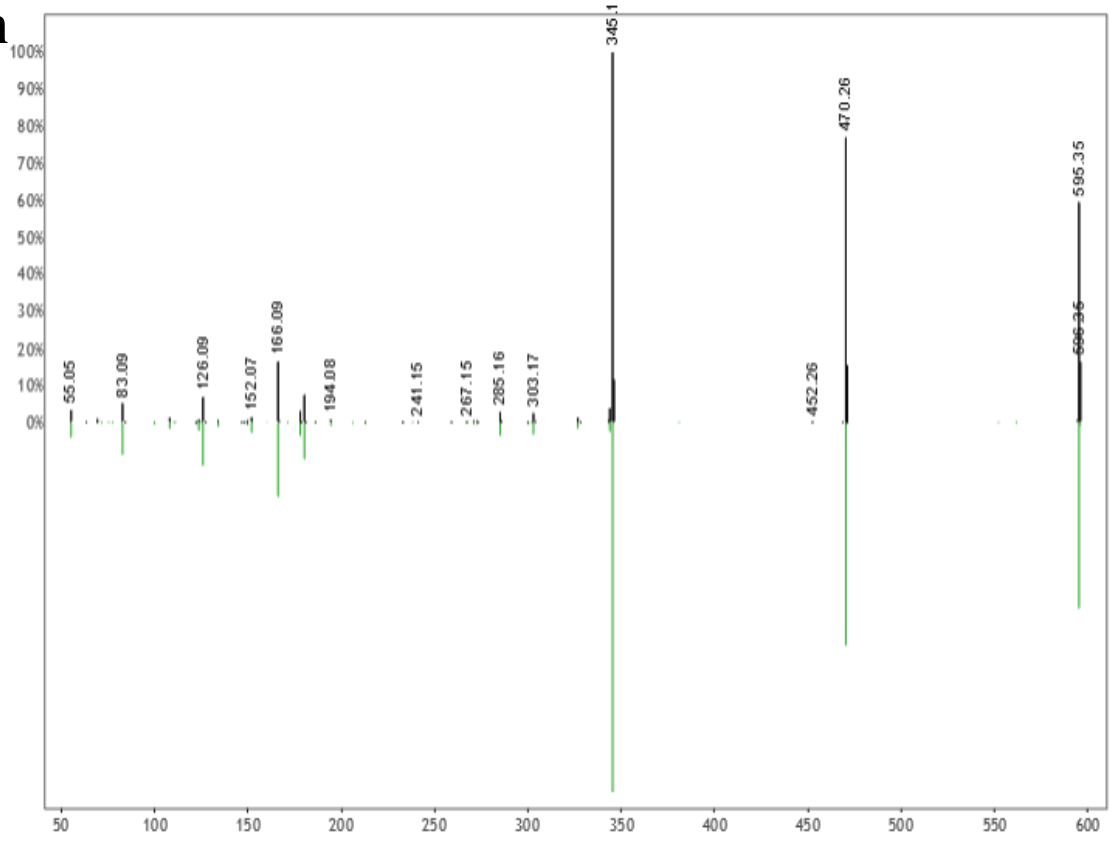
bl


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
59.05
95.09
109.10
145.10
160.09
197.13
227.12
257.13
286.13
299.1
504.36
586.28
50
100
150
200
250
300
350
400
450
500
550

bm


343.1
591.32
596.52
59.61
89.06
124.08
136.08
164.07
211.12
225.14
269.13
283.14
299.14
394.21
408.23
438.20
466.23
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
100
150
200
250
300
350
400
450
500
550
600

bn


345.1
470.26
595.35
55.05
83.09
126.09
152.07
166.09
194.08
241.15
267.15
285.16
303.17
452.26
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50
100
150
200
250
300
350
400
450
500
550
600

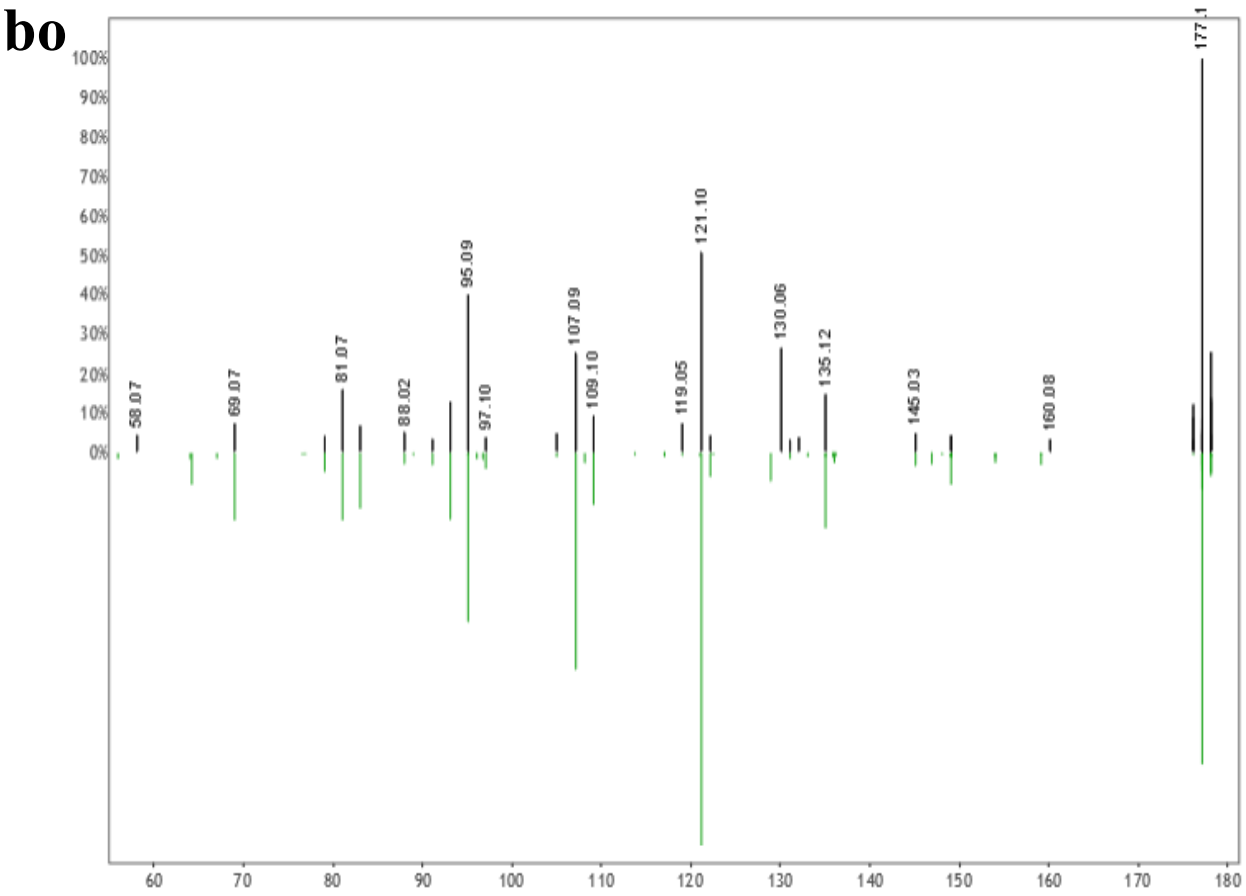
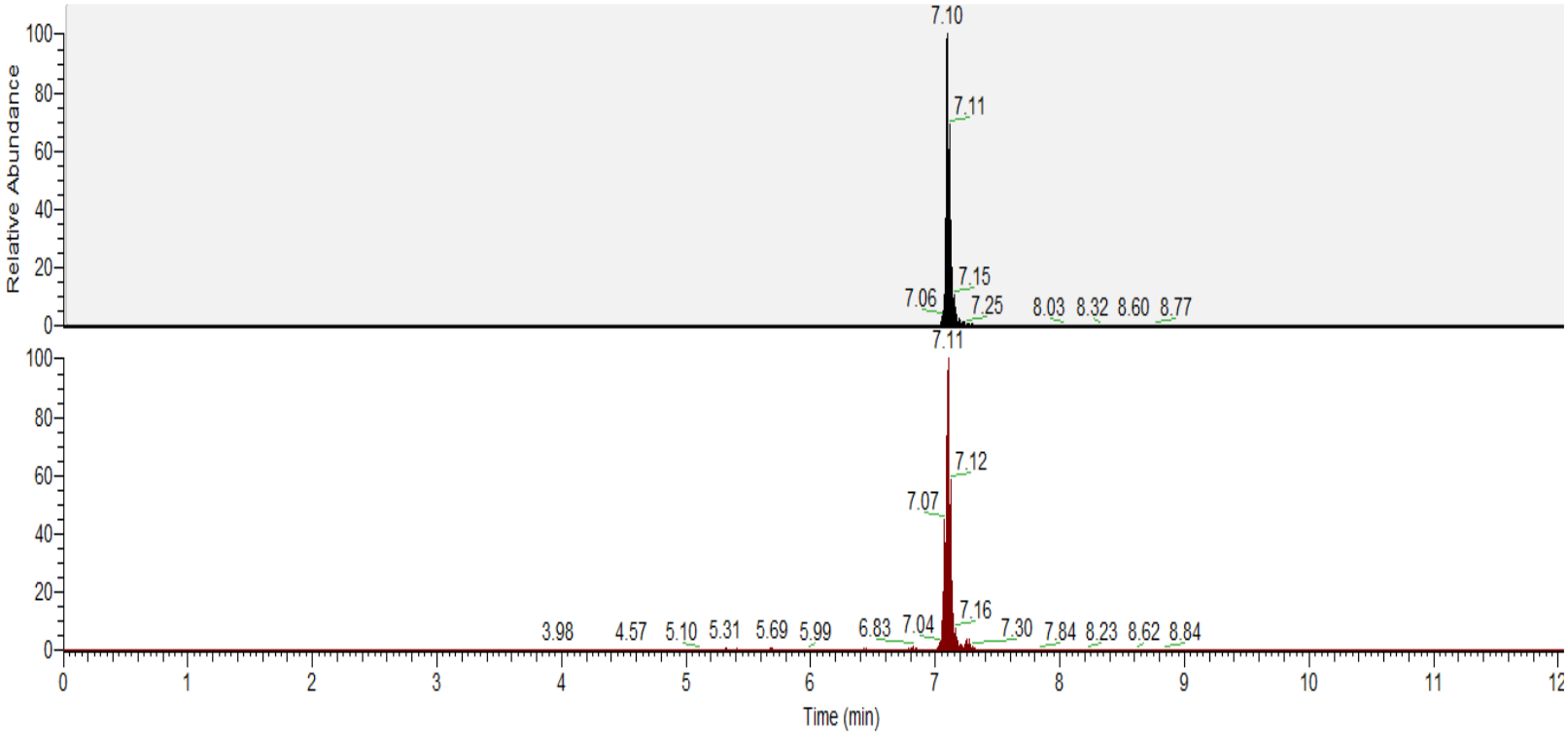


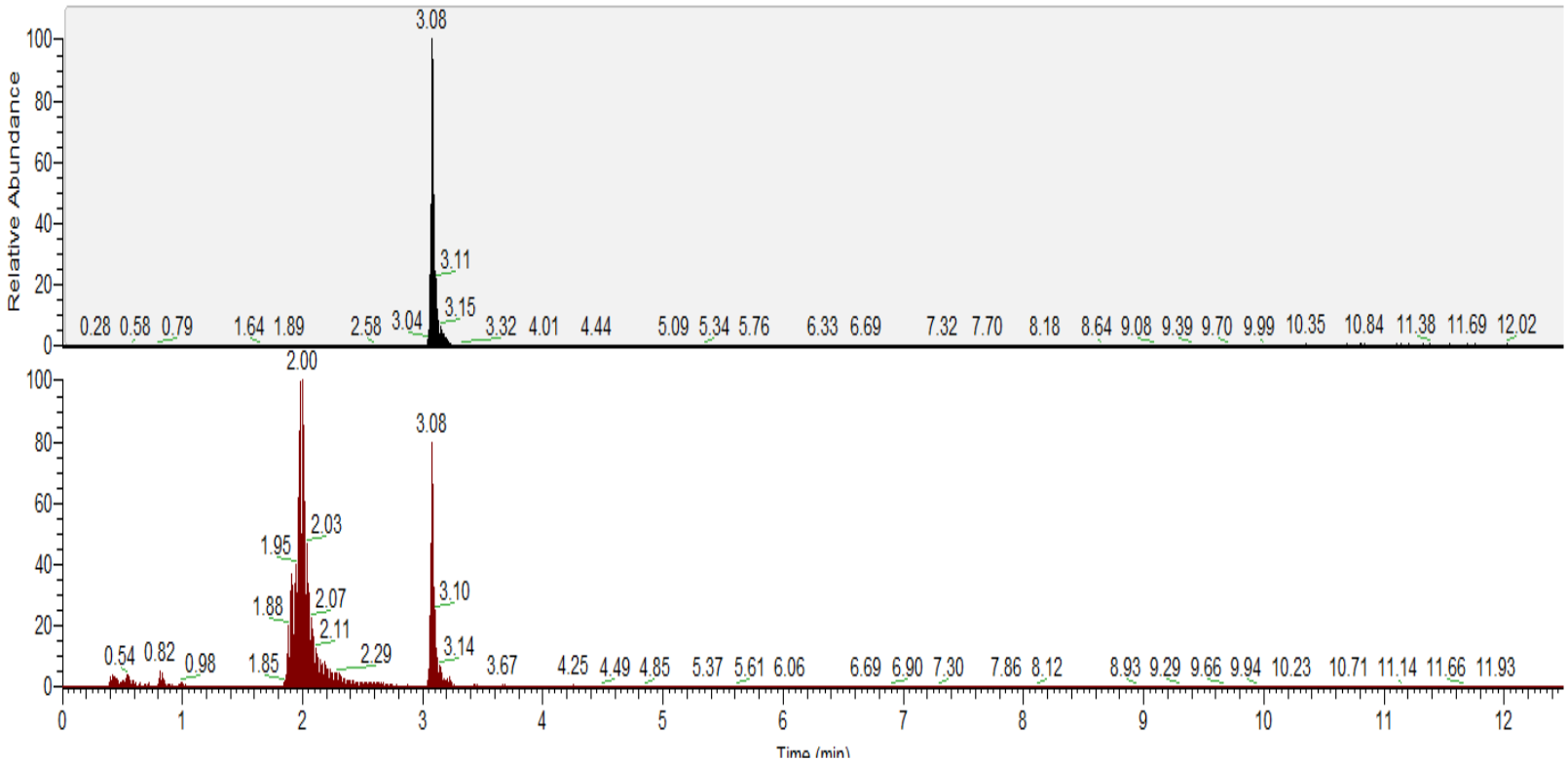
Figure S2.3: Mirror plots of shared human fecal metabolites.
Top black spectra represent experimental MS2 for the respective metabolite, while bottom green spectra represent library reference MS2 spectra. ***a****, Hypoxanthine;* ***b****, Nicotinamide N-oxide;* ***c****, 3-methyl-2-oxindole (3-Methyloxindole);* ***d****, Hyocholic acid;* ***e****, Gly-Val (Glycylvaline);* ***f****, 3-Hydroxy-4-methoxycinnamic acid (Isoferulic acid);* ***g****, Paraxanthine;* ***h****, Phe-Pro (Phenylalanylproline);* ***i****, trans-Ferulic acid;* ***j****, Loliolide;* ***k****, N-Acetyl-D-mannosamine;* ***l****, N-acetyl-L-Phenylalanine;* ***m****, Thr-Pro (Threonylproline);* ***n****, Val-Val (Valylvaline);* ***o****, Abrine;* ***p****, Pantothenic acid;* ***q****, PyroGlu-Pro (Pyroglutamylproline);* ***r****, Ile-Pro (Isoleucylproline);* ***s****, Val-Ile (Valylisoleucine);* ***t****, cis-9-Hexadecenoic acid (Palmitoleic acid);* ***u****, Palmitelaidic acid;* ***v****, Gly-Tyr (Glycyltyrosine);* ***w****, Biotin;* ***x****, Leu-Leu (Leucylleucine);* ***y****, Lenticin;* ***z****, Val-Met (Valylmethionine);* ***aa****, Ser-Phe (Serylphenylalanine);* ***ab****, L-Saccharopine;* ***ac****, Conjugated linoleic Acid (10E,12Z);* ***ad****, Conjugated linoleic acid (9E,11E);* ***ae****, Thr-Phe (Threonylphenylalanine);* ***af****, Pro-Arg (Prolylarginine);* ***ag****, 9-OxoOTrE;* ***ah****, Phe-Leu (Phenylalanylleucine);* ***ai****, Leu-Phe (Leucylphenylalanine);* ***aj****, Xanthosine;* ***ak****, Arg-Ile (Arginylisoleucine);* ***al****, N-Tetracosenoyl-4-sphingenine;* ***am****, N-Acetylmuramic Acid;* ***an****, Tyr-Leu (Tyrosylleucine);* ***ao****, Phe-Met (Phenylalanylmethionine);* ***ap****, Ile-Gly-Ile (Isoleucylglycylisoleucine);* ***aq****, cis-11,14-Eicosadienoic acid;* ***ar****, Val-Trp (Valyltryptophan);* ***as****, Fructoselysine;* ***at****, Myristoleic acid;* ***au****, N-Palmitoylglycine;* ***av****, Ile-Trp (Iseoleucyltryptophan);* ***aw****, Lithocholic acid;* ***ax****, Leucine enkephalin;* ***ay****, 13-Docosenamide, (Z)- (Erucamide);* ***az****, Phe-Trp (Phenylalanyltryptophan);* ***ba****, Ile-Val-Lys (Isoleucylvalyllysine);* ***bb****, 3-Hydroxydodecanoic*

acid; ***bc****, Cholesterol;* ***bd****, 6R)-2-(hydroxymethyl)-6-((3R,5R,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)heptanoic acid;* ***be****, Cholic acid;* ***bf****, Octadecanamide;* ***bg****, (R)-4-((3R,5S,8R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-7,12-dioxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid;* ***bh****, Glycoursodeoxycholic acid;* ***bi****, Oleanolic acid;* ***bj****, Glycocholic acid;* ***bk****, Enoxolone;* ***bl****, Bilirubin;* ***bm****, Urobilin;* ***bn****, Stercobilin;* ***bo****, 2-Butanone, 4-(2,6,6-trimethyl-2-cyclohexen-1-yl).*

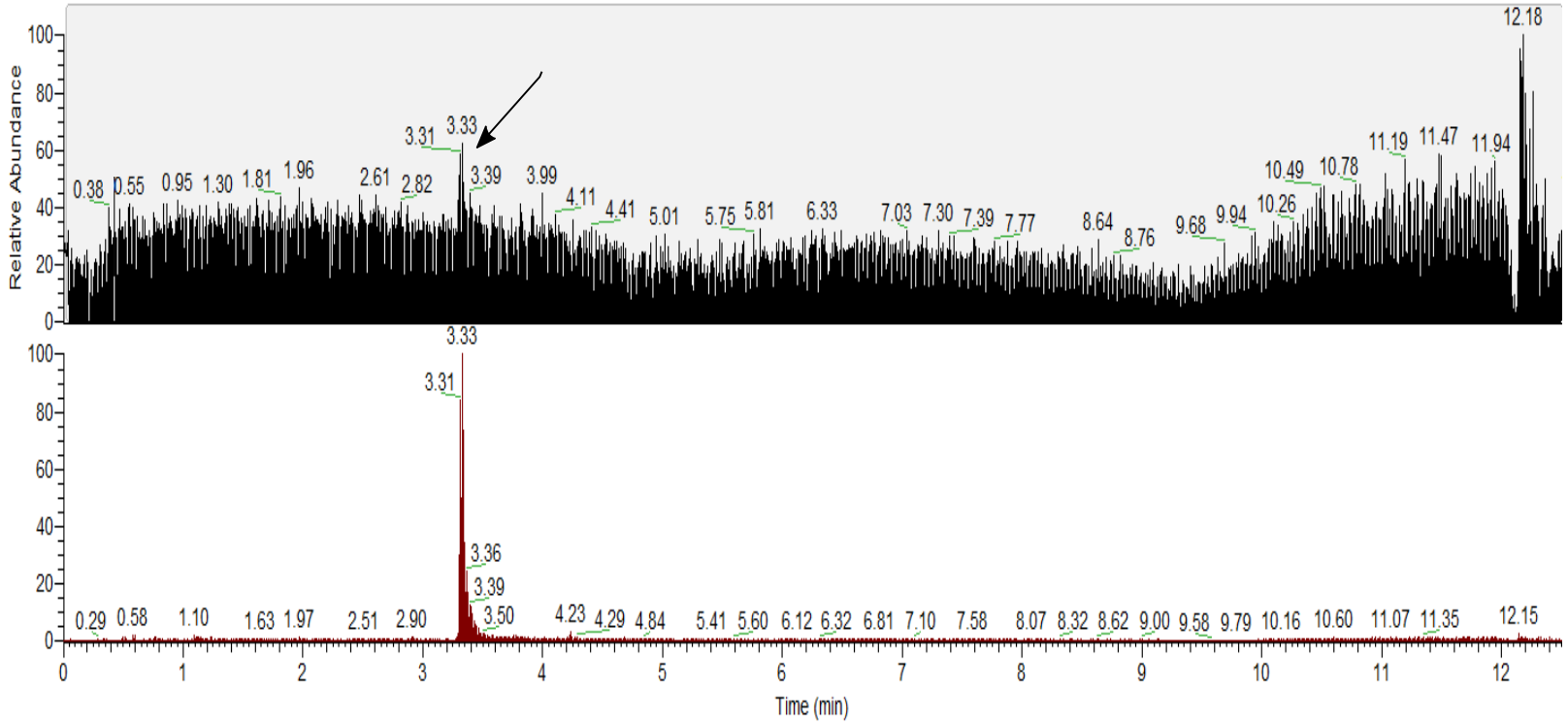
a


Relative Abundance
Time (min)
7.10
7.11
7.15
7.06
7.25
8.03
8.32
8.60
8.77
7.11
7.12
7.07
7.16
3.98
4.57
5.10
5.31
5.69
5.99
6.83
7.04
7.30
7.84
8.23
8.62
8.84

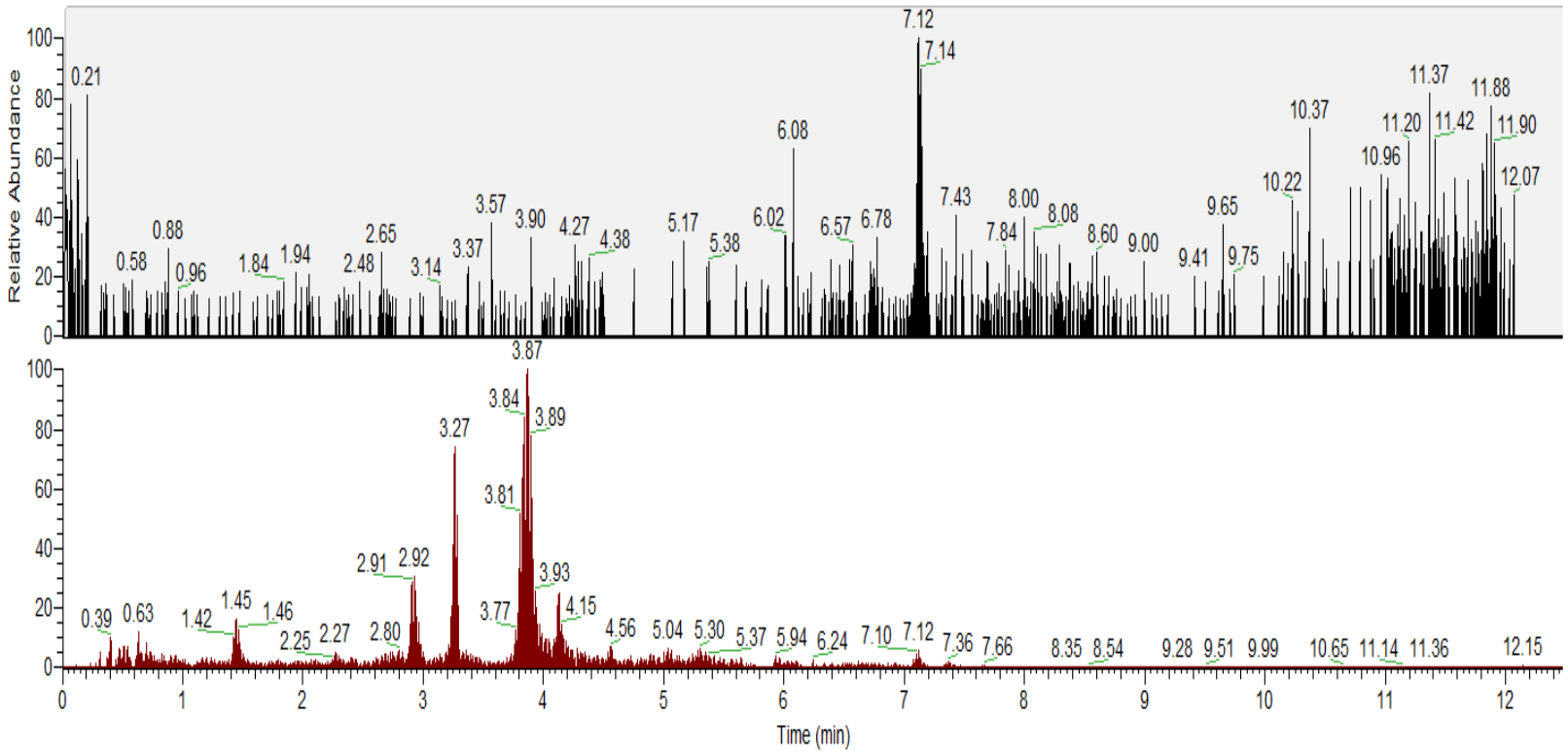
b


Relative Abundance
Time (min)
3.08
3.11
3.15
2.00
3.08
2.03
1.95
2.07
1.88
2.11
3.10
3.14
2.29
1.85
0.54
0.82
0.98

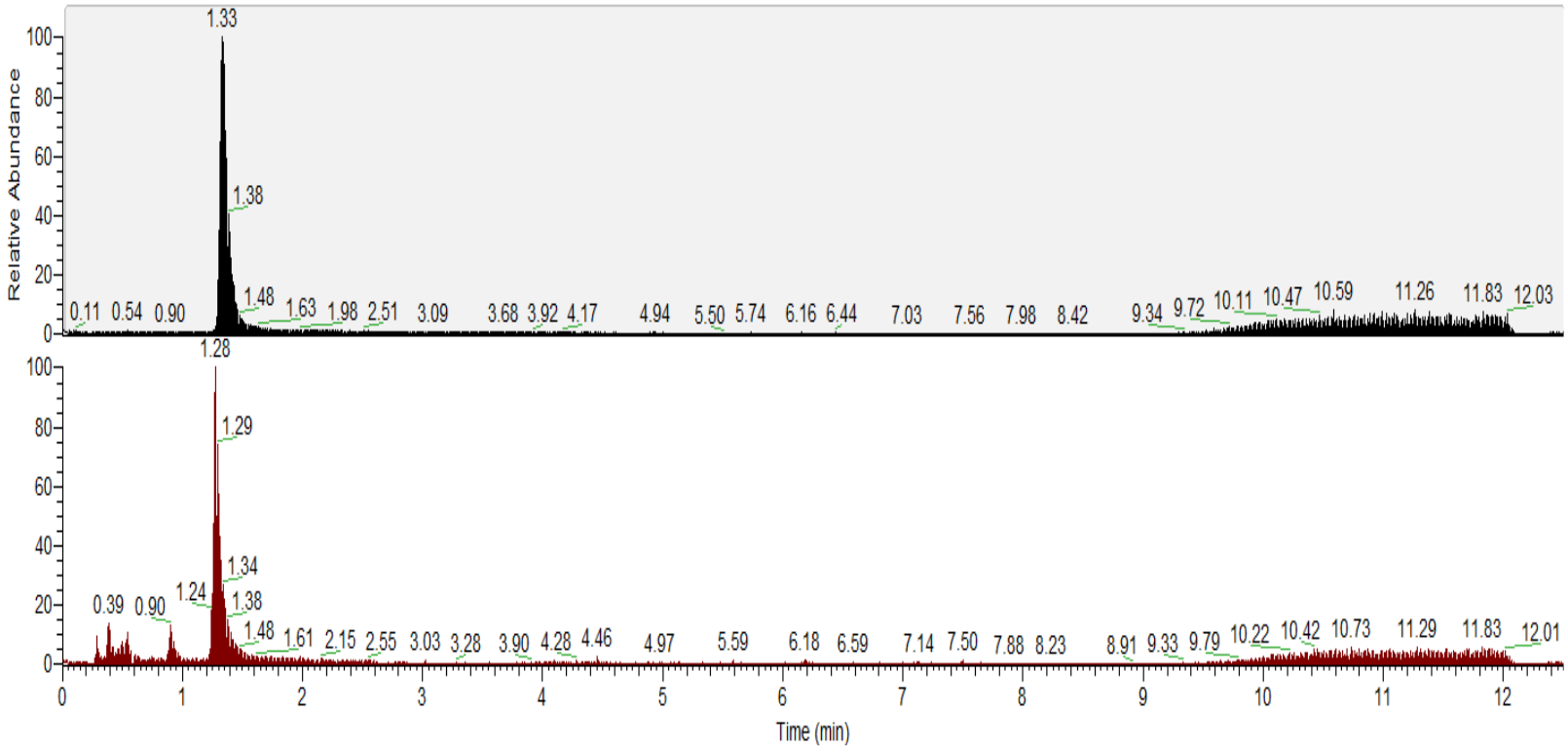
c


Relative Abundance
Time (min)
0.38 0.55 0.95 1.30 1.81 1.96 2.61 2.82 3.31 3.33 3.39 3.99 4.11 4.41 5.01 5.75 5.81 6.33 7.03 7.30 7.39 7.77 8.64 8.76 9.68 9.94 10.26 10.49 10.78 11.19 11.47 11.94 12.18
0.29 0.58 1.10 1.63 1.97 2.51 2.90 3.31 3.33 3.36 3.39 3.50 4.23 4.29 4.84 5.41 5.60 6.12 6.32 6.81 7.10 7.58 8.07 8.32 8.62 9.00 9.58 9.79 10.16 10.60 11.07 11.35 12.15

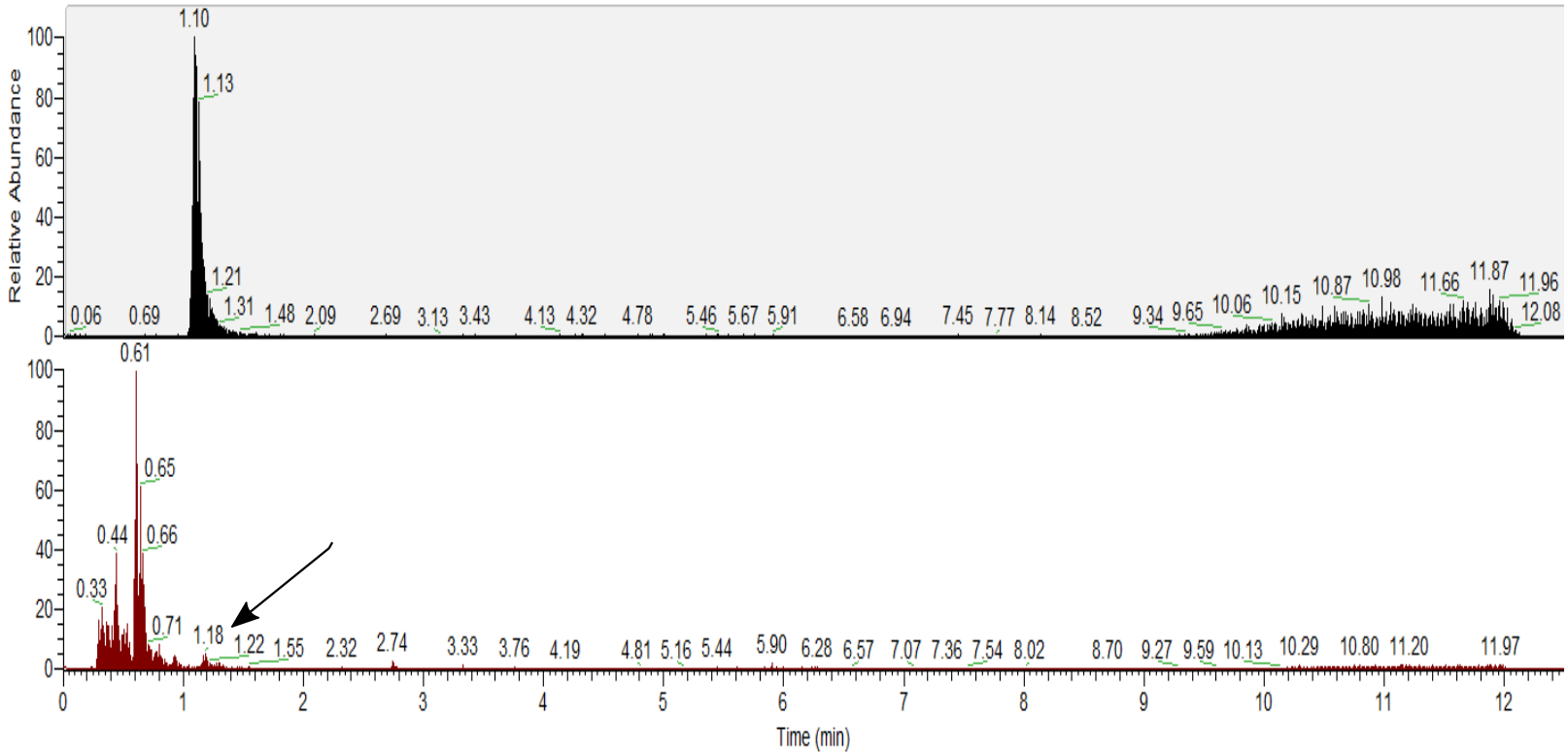
d


Relative Abundance
Time (min)
0.21
0.58
0.88
0.96
1.84
1.94
2.48
2.65
3.14
3.37
3.57
3.90
4.27
4.38
5.17
5.38
6.02
6.08
6.57
6.78
7.12
7.14
7.43
7.84
8.00
8.08
8.60
9.00
9.41
9.65
9.75
10.22
10.37
10.96
11.20
11.37
11.42
11.88
11.90
12.07
0.39
0.63
1.42
1.45
1.46
2.25
2.27
2.80
2.91
2.92
3.27
3.77
3.81
3.84
3.87
3.89
3.93
4.15
4.56
5.04
5.30
5.37
5.94
6.24
7.10
7.12
7.36
7.66
8.35
8.54
9.28
9.51
9.99
10.65
11.14
11.36
12.15

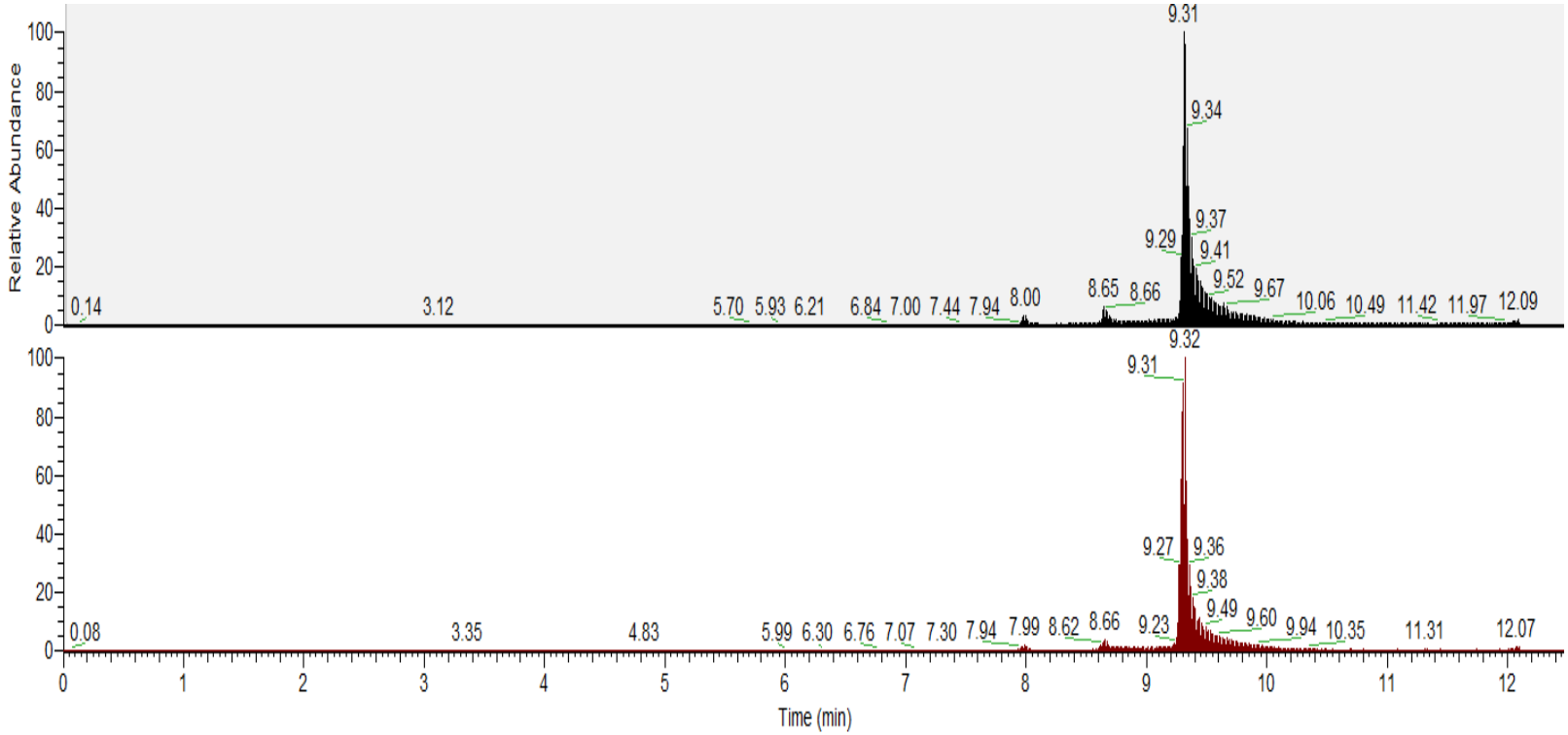
e



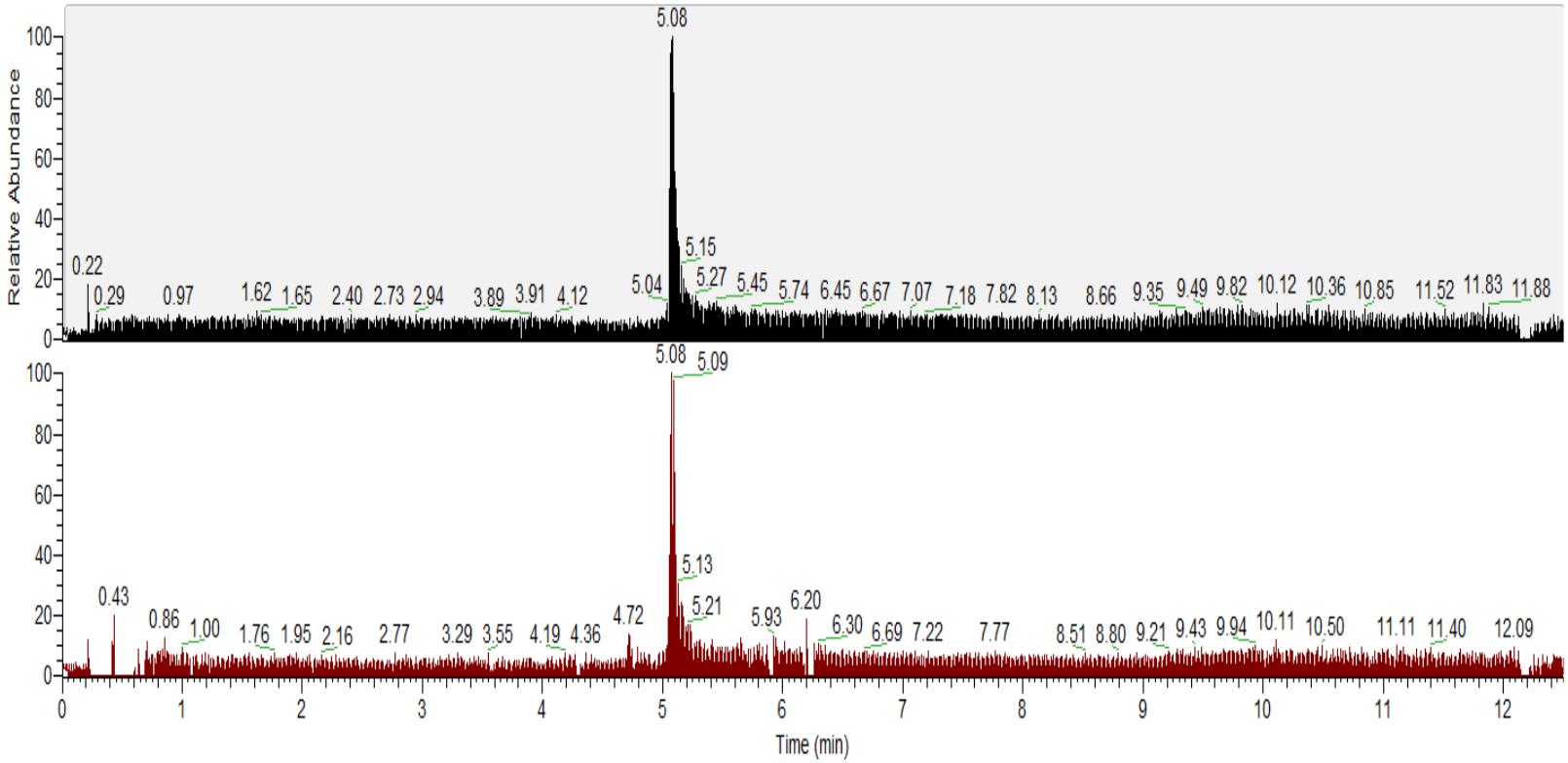
f


Relative Abundance
Time (min)
1.10
1.13
1.21
1.31
1.48
2.09
2.69
3.13
3.43
4.13
4.32
4.78
5.46
5.67
5.91
6.58
6.94
7.45
7.77
8.14
8.52
9.34
9.65
10.06
10.15
10.87
10.98
11.66
11.87
11.96
12.08
0.06
0.69
0.61
0.65
0.44
0.66
0.33
0.71
1.18
1.22
1.55
2.32
2.74
3.33
3.76
4.19
4.81
5.16
5.44
5.90
6.28
6.57
7.07
7.36
7.54
8.02
8.70
9.27
9.59
10.13
10.29
10.80
11.20
11.97

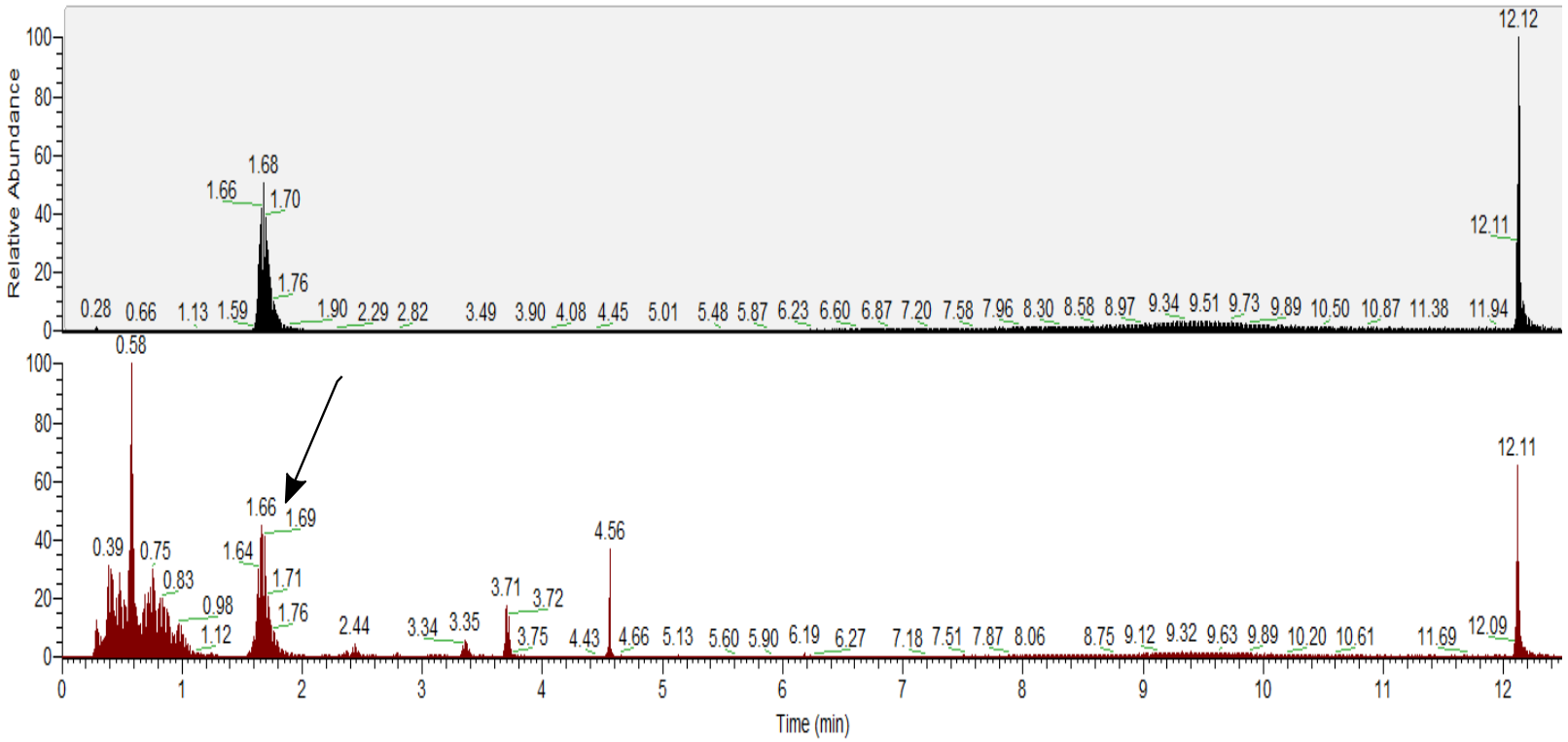
g


Relative Abundance
Time (min)
9.31
9.34
9.37
9.29
9.41
9.52
9.67
10.06
10.49
11.42
11.97
12.09
0.14
3.12
5.70
5.93
6.21
6.84
7.00
7.44
7.94
8.00
8.65
8.66
9.32
9.31
9.27
9.36
9.38
9.49
9.60
9.94
10.35
11.31
12.07
0.08
3.35
4.83
5.99
6.30
6.76
7.07
7.30
7.94
7.99
8.62
8.66
9.23

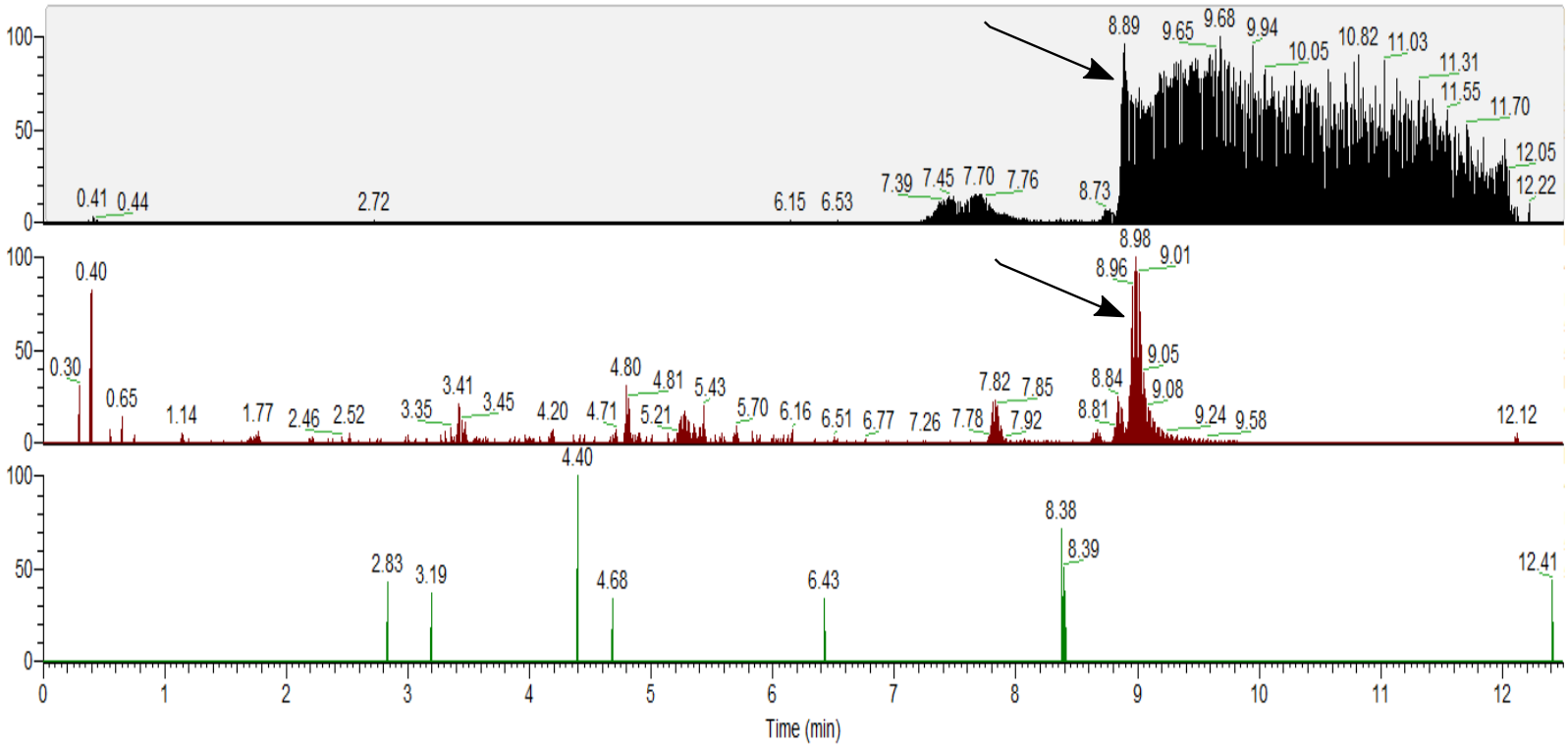
h


Relative Abundance
Time (min)
5.08
5.15
5.27
5.45
5.74
6.45
6.67
7.07
7.18
7.82
8.13
8.66
9.35
9.49
9.82
10.12
10.36
10.85
11.52
11.83
11.88
0.22
0.29
0.97
1.62
1.65
2.40
2.73
2.94
3.89
3.91
4.12
5.04
5.08
5.09
5.13
5.21
5.93
6.20
6.30
6.69
7.22
7.77
8.51
8.80
9.21
9.43
9.94
10.11
10.50
11.11
11.40
12.09
0.43
0.86
1.00
1.76
1.95
2.16
2.77
3.29
3.55
4.19
4.36
4.72

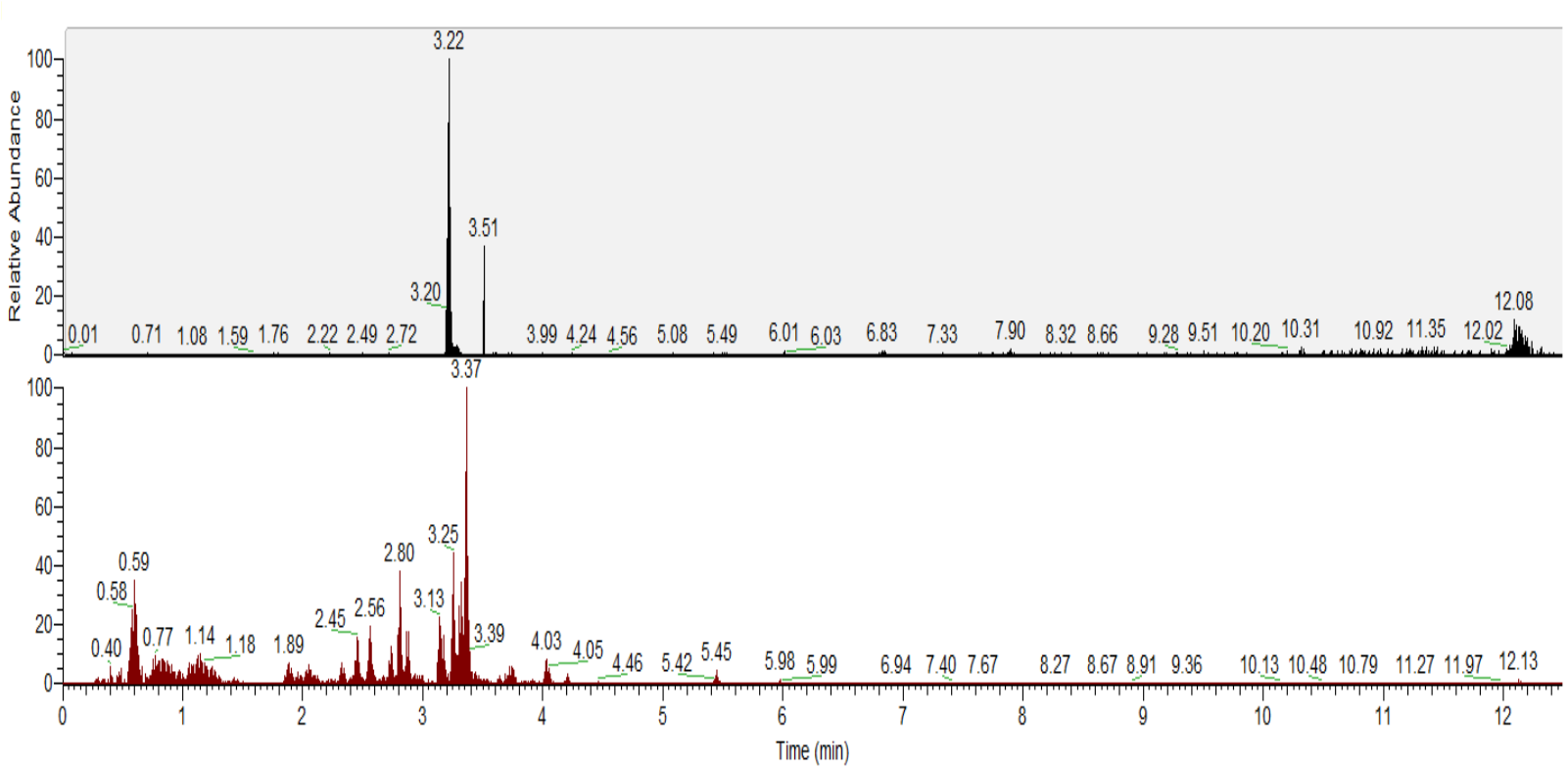
i


Relative Abundance
Time (min)
12.12
1.68
1.66
1.70
12.11
1.76
0.28 0.66 1.13 1.59 1.90 2.29 2.82 3.49 3.90 4.08 4.45 5.01 5.48 5.87 6.23 6.60 6.87 7.20 7.58 7.96 8.30 8.58 8.97 9.34 9.51 9.73 9.89 10.50 10.87 11.38 11.94
0.58
12.11
1.66
1.69
4.56
0.39
0.75
1.64
0.83
1.71
3.71
3.72
0.98
1.76
1.12
2.44
3.34
3.35
3.75
4.43
4.66
5.13
5.60
5.90
6.19
6.27
7.18
7.51
7.87
8.06
8.75
9.12
9.32
9.63
9.89
10.20
10.61
11.69
12.09

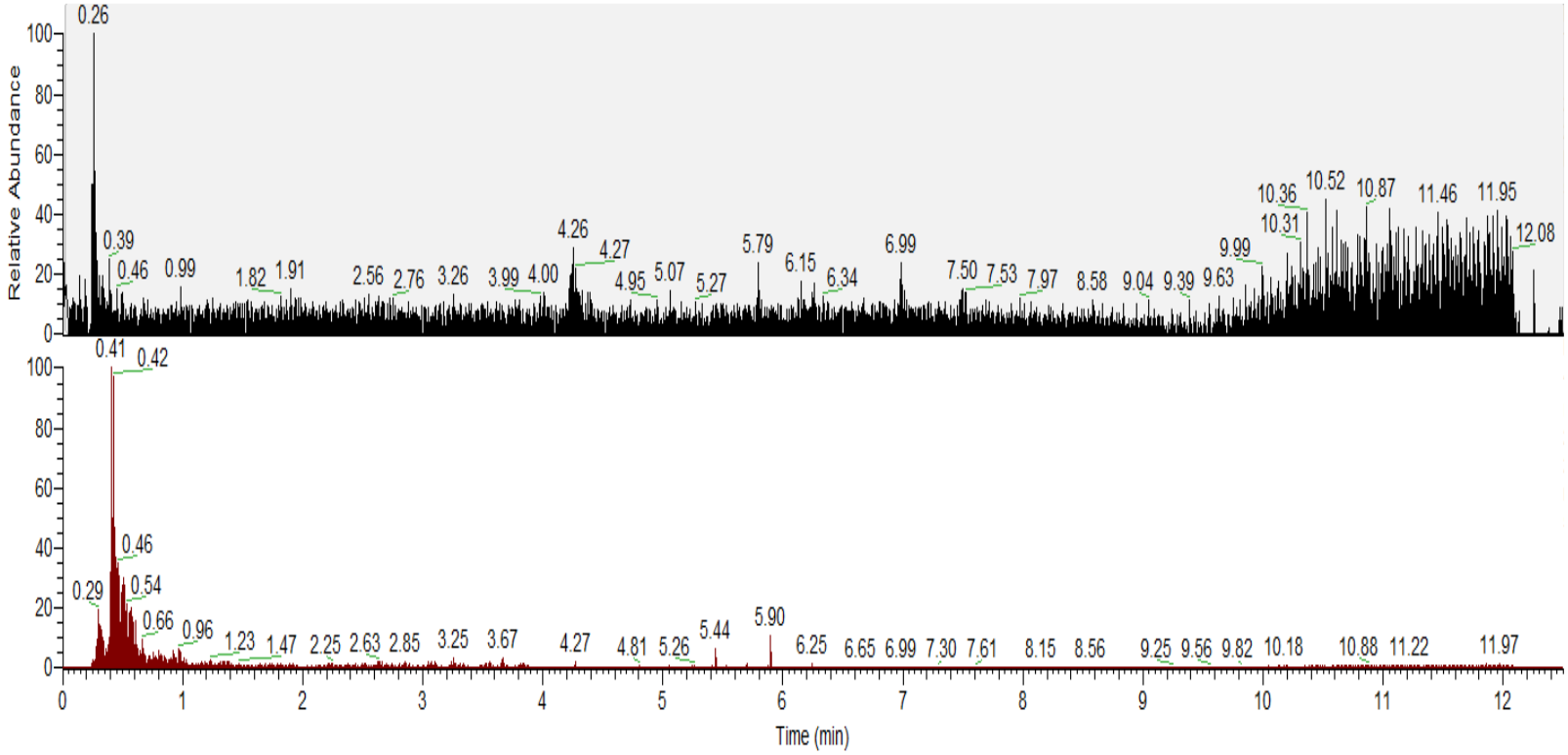
j


100
50
0
0.41 0.44
2.72
6.15 6.53
7.39 7.45 7.70 7.76
8.73
8.89 9.65 9.68 9.94 10.05 10.82 11.03 11.31 11.55 11.70 12.05 12.22
0.30 0.40 0.65 1.14 1.77 2.46 2.52 3.35 3.41 3.45 4.20 4.71 4.80 4.81 5.21 5.43 5.70 6.16 6.51 6.77 7.26 7.78 7.82 7.85 7.92 8.81 8.84 8.96 8.98 9.01 9.05 9.08 9.24 9.58 12.12
2.83 3.19 4.40 4.68 6.43 8.38 8.39 12.41
0 1 2 3 4 5 6 7 8 9 10 11 12
Time (min)

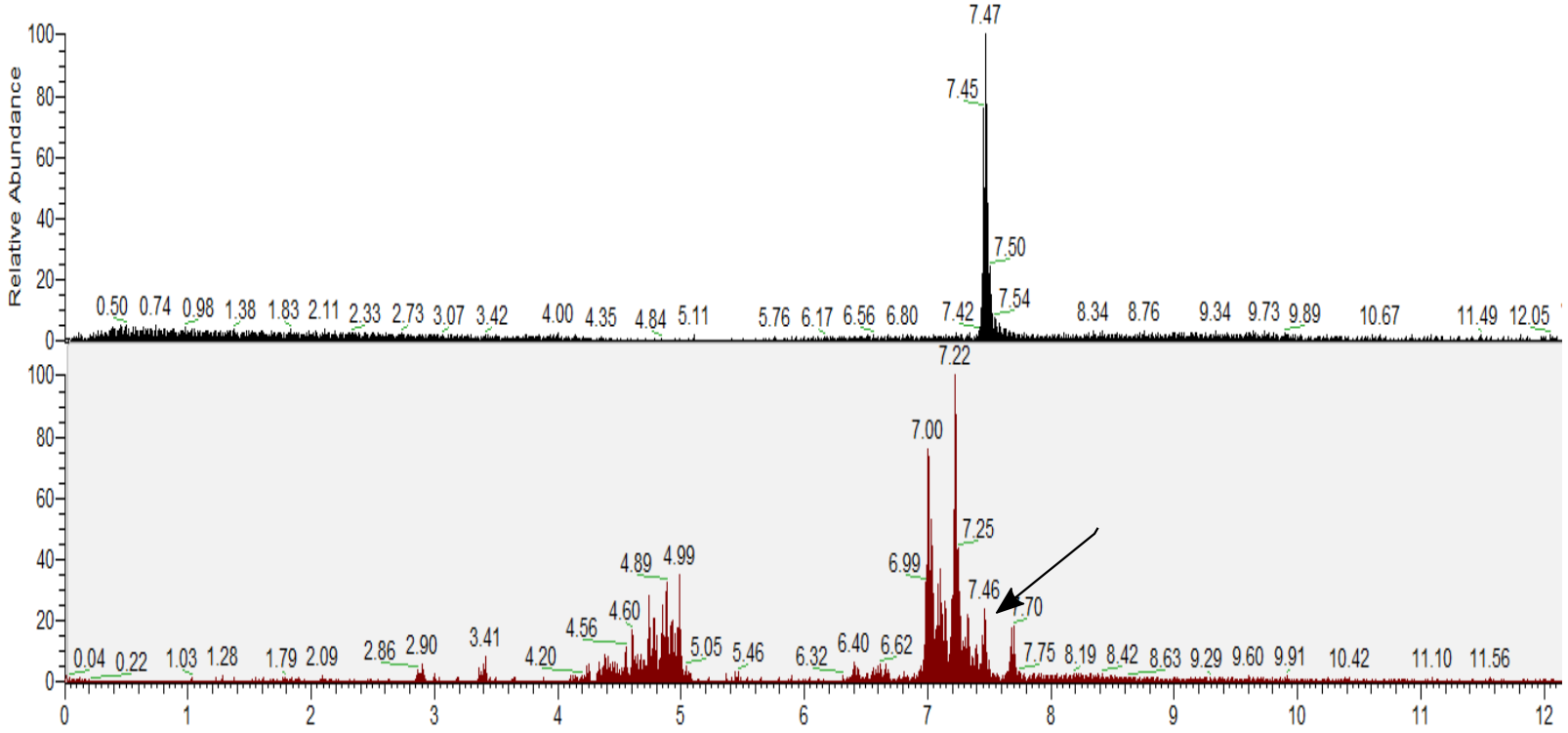
k


Relative Abundance
Time (min)
3.22
3.51
3.20
12.08
3.37
3.25
2.80
0.59
0.58
3.13
2.45
2.56
3.39
0.40
0.77
1.14
1.18
1.89
4.03
4.05
4.46
5.42
5.45
5.98
5.99
6.94
7.40
7.67
8.27
8.67
8.91
9.36
10.13
10.48
10.79
11.27
11.97
12.13

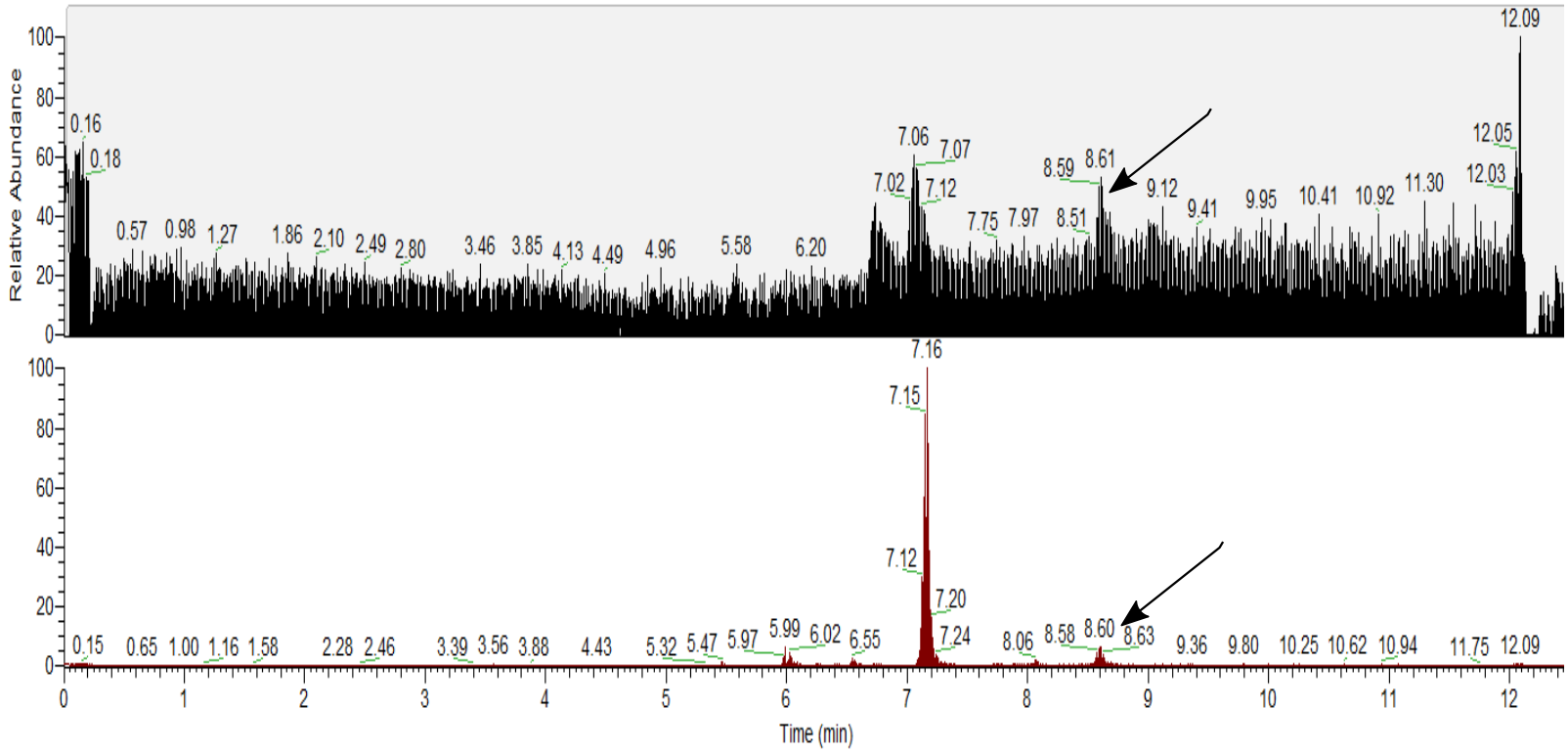
I


Relative Abundance
Time (min)
0.26
0.39
0.46
0.99
1.82
1.91
2.56
2.76
3.26
3.99
4.00
4.26
4.27
4.95
5.07
5.27
5.79
6.15
6.34
6.99
7.50
7.53
7.97
8.58
9.04
9.39
9.63
9.99
10.31
10.36
10.52
10.87
11.46
11.95
12.08
0.29
0.41
0.42
0.46
0.54
0.66
0.96
1.23
1.47
2.25
2.63
2.85
3.25
3.67
4.27
4.81
5.26
5.44
5.90
6.25
6.65
6.99
7.30
7.61
8.15
8.56
9.25
9.56
9.82
10.18
10.88
11.22
11.97

m


Relative Abundance
7.47
7.45
7.50
7.54
0.50 0.74 0.98 1.38 1.83 2.11 2.33 2.73 3.07 3.42 4.00 4.35 4.84 5.11 5.76 6.17 6.56 6.80 7.42 8.34 8.76 9.34 9.73 9.89 10.67 11.49 12.05
7.22
7.00
7.25
4.89 4.99
6.99
7.46
4.60
4.56
7.70
0.04 0.22 1.03 1.28 1.79 2.09 2.86 2.90 3.41 4.20 5.05 5.46 6.32 6.40 6.62 7.75 8.19 8.42 8.63 9.29 9.60 9.91 10.42 11.10 11.56
0 1 2 3 4 5 6 7 8 9 10 11 12

n


Relative Abundance
Time (min)
0.16
0.18
0.57
0.98
1.27
1.86
2.10
2.49
2.80
3.46
3.85
4.13
4.49
4.96
5.58
6.20
7.02
7.06
7.07
7.12
7.75
7.97
8.51
8.59
8.61
9.12
9.41
9.95
10.41
10.92
11.30
12.03
12.05
12.09
7.16
7.15
7.12
7.20
0.15
0.65
1.00
1.16
1.58
2.28
2.46
3.39
3.56
3.88
4.43
5.32
5.47
5.97
5.99
6.02
6.55
7.24
8.06
8.58
8.60
8.63
9.36
9.80
10.25
10.62
10.94
11.75
12.09

o

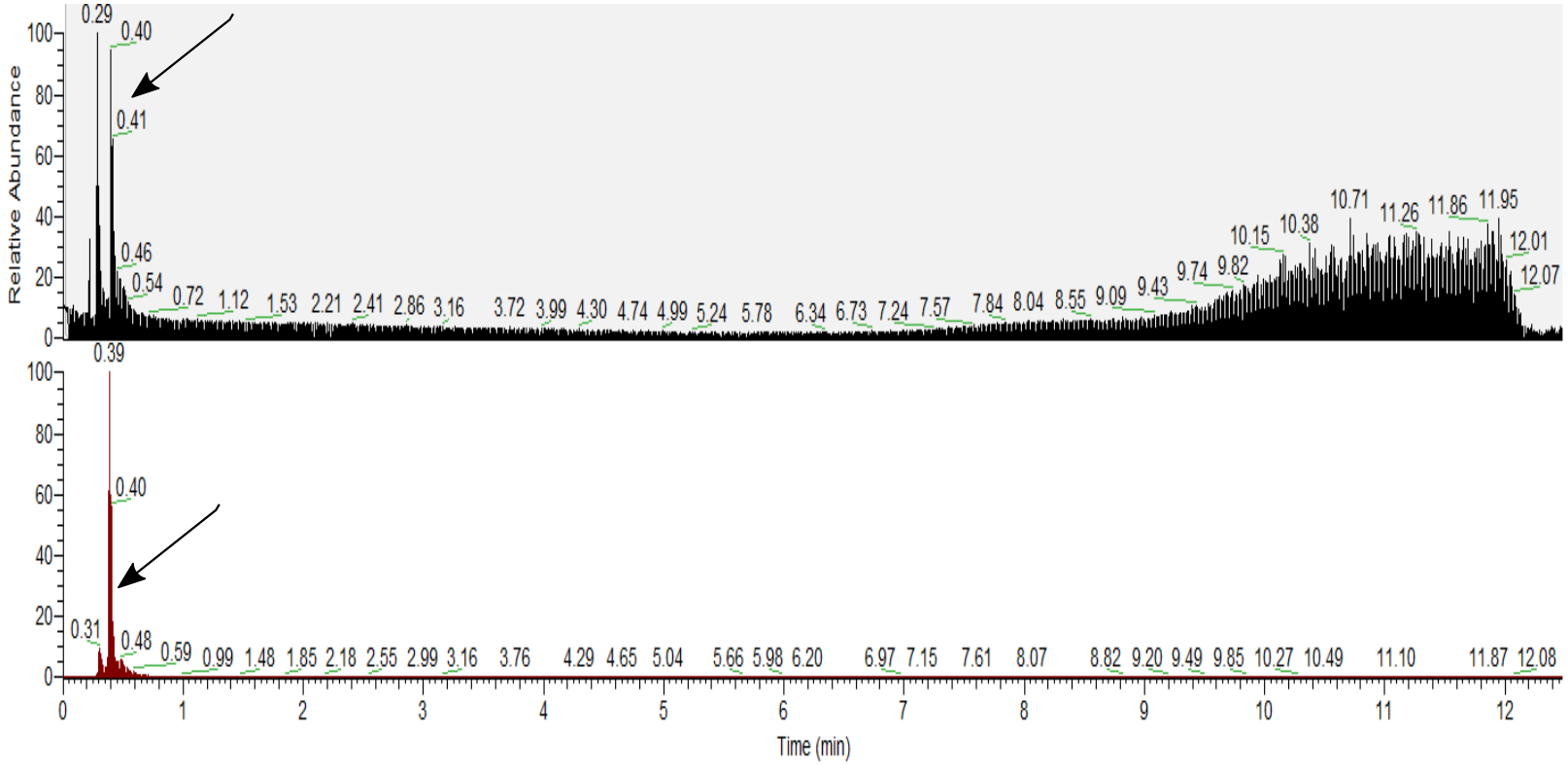

Relative Abundance
Time (min)
0.29
0.40
0.41
0.46
0.54
0.72
1.12
1.53
2.21
2.41
2.86
3.16
3.72
3.99
4.30
4.74
4.99
5.24
5.78
6.34
6.73
7.24
7.57
7.84
8.04
8.55
9.09
9.43
9.74
9.82
10.15
10.38
10.71
11.26
11.86
11.95
12.01
12.07
0.39
0.40
0.31
0.48
0.59
0.99
1.48
1.85
2.18
2.55
2.99
3.16
3.76
4.29
4.65
5.04
5.66
5.98
6.20
6.97
7.15
7.61
8.07
8.82
9.20
9.49
9.85
10.27
10.49
11.10
11.87
12.08

Figure S2.4. Authentic standards validating sample annotations.
Standards are in black, representative samples are in red, blanks (when displayed) are in green. ***a****, Enoxolone;* ***b****, N-acetyl-L-phenylalanine;* ***c****, trans-Ferulic acid;* ***d****, Lithocholic acid;* ***e****, Paraxanthine;* ***f****, L-Abrine;* ***g****, 13-Docosenamide, (Z)-;* ***h****, Hyocholic acid;* ***i****, Lenticin;* ***j****, Leucine enkephalin;* ***k****, L-Saccharopine;* ***l****, N-palmitoylglycine;* ***m****, Nicotinamide N-oxide.*

Supplementary Tables 2.1-2.4

Table S2.1: Metadata for individual samples.

SampleID	Population	Industrialization Category	Sex	Age
NO01	Norman	Urban Industrial	M	23
NO02	Norman	Urban Industrial	F	37
NO03	Norman	Urban Industrial	M	40
NO04	Norman	Urban Industrial	M	26
NO05	Norman	Urban Industrial	M	28
NO06	Norman	Urban Industrial	M	28
NO07	Norman	Urban Industrial	F	32
NO08	Norman	Urban Industrial	F	32
NO09	Norman	Urban Industrial	F	34
NO10	Norman	Urban Industrial	M	41
NO11	Norman	Urban Industrial	M	26
NO12	Norman	Urban Industrial	F	27
NO13	Norman	Urban Industrial	M	35
NO19	Norman	Urban Industrial	F	32
NO20	Norman	Urban Industrial	M	26
NO21	Norman	Urban Industrial	M	23
NO22	Norman	Urban Industrial	M	26
NO23	Norman	Urban Industrial	F	26
GU1	Guayabo	Rural Industrial	F	52
GU2	Guayabo	Rural Industrial	F	19
GU4	Guayabo	Rural Industrial	F	40
GU6	Guayabo	Rural Industrial	NA	6
GU7	Guayabo	Rural Industrial	NA	9
GU10	Guayabo	Rural Industrial	F	58
GU11	Guayabo	Rural Industrial	NA	7
GU12	Guayabo	Rural Industrial	NA	16
GU13	Guayabo	Rural Industrial	F	41
GU17	Guayabo	Rural Industrial	F	51
GU19	Guayabo	Rural Industrial	F	24
GU20	Guayabo	Rural Industrial	F	63
1TM	Tambo de Mora	Rural Industrial	F	61

2TM	Tambo de Mora	Rural Industrial	F	40
3TM	Tambo de Mora	Rural Industrial	NA	5
4TM	Tambo de Mora	Rural Industrial	F	40
6TM	Tambo de Mora	Rural Industrial	F	31
10TM	Tambo de Mora	Rural Industrial	NA	8
11TM	Tambo de Mora	Rural Industrial	M	39
14TM	Tambo de Mora	Rural Industrial	F	77
16TM	Tambo de Mora	Rural Industrial	F	38
17TM	Tambo de Mora	Rural Industrial	NA	7
18TM	Tambo de Mora	Rural Industrial	NA	13
26TM	Tambo de Mora	Rural Industrial	F	36
27TM	Tambo de Mora	Rural Industrial	NA	13
31TM	Tambo de Mora	Rural Industrial	NA	28
TM10_01	Burkina Faso	Rural Traditional	M	55
TM13_01	Burkina Faso	Rural Traditional	M	53
TM23_02	Burkina Faso	Rural Traditional	F	51
TM01_01	Burkina Faso	Rural Traditional	M	55
TM09_02	Burkina Faso	Rural Traditional	F	32
TM11_04	Burkina Faso	Rural Traditional	F	40
TM17_02	Burkina Faso	Rural Traditional	F	35

TM20_03	Burkina Faso	Rural Traditional	F	37
TM22_03	Burkina Faso	Rural Traditional	M	29
TM25_03	Burkina Faso	Rural Traditional	F	38
TM29_01	Burkina Faso	Rural Traditional	M	73
HCO01	Tunapuco	Rural Traditional	F	36
HCO03	Tunapuco	Rural Traditional	M	6
HCO09	Tunapuco	Rural Traditional	NA	13
HCO10	Tunapuco	Rural Traditional	M	10
HCO11	Tunapuco	Rural Traditional	M	36
HCO12	Tunapuco	Rural Traditional	F	35
HCO13	Tunapuco	Rural Traditional	F	9
HCO14	Tunapuco	Rural Traditional	F	34
HCO15	Tunapuco	Rural Traditional	F	63
HCO16	Tunapuco	Rural Traditional	NA	11
HCO17	Tunapuco	Rural Traditional	F	7
HCO18	Tunapuco	Rural Traditional	M	11
HCO21	Tunapuco	Rural Traditional	M	10
HCO41	Tunapuco	Rural Traditional	NA	54
HCO53	Tunapuco	Rural Traditional	F	44
HCO61	Tunapuco	Rural Traditional	F	20
HCO63	Tunapuco	Rural Traditional	F	6
HCO66	Tunapuco	Rural Traditional	M	11
HCO67	Tunapuco	Rural Traditional	F	26
HCO68	Tunapuco	Rural Traditional	M	7
HCO69	Tunapuco	Rural Traditional	NA	9
HCO70	Tunapuco	Rural Traditional	F	40
HCO72	Tunapuco	Rural Traditional	F	5
HCO74	Tunapuco	Rural Traditional	F	36
SM01	Matses	Isolated Traditional	M	30
SM02	Matses	Isolated Traditional	F	25
SM03	Matses	Isolated Traditional	M	10

SM10	Matses	Isolated Traditional	F	6
SM23	Matses	Isolated Traditional	M	7
SM28	Matses	Isolated Traditional	F	52
SM29	Matses	Isolated Traditional	F	50
SM33	Matses	Isolated Traditional	F	5
SM37	Matses	Isolated Traditional	M	12
SM39	Matses	Isolated Traditional	F	40
SM41	Matses	Isolated Traditional	M	6

Table S2.2: The shared human fecal metabolome.
Results are derived from both gap-filled and non-gap-filled analyses. All annotations had a mass difference of 0 to library reference. Listed m/z and RT values are derived from gap-filled analyses, but all annotations were present in both gap-filled and non-gap-filled analyses. Bolded rows represent features detected in all ReDU datasets. For predicted ClassyFire[104] classes column, percentages represent a chemical classification score reporting the percentage of nodes assigned to the specified chemical class within the molecular network cluster.

Compound Name	***m/z***	**RT (min)**	**Cosine Score**	**Detected in 6-sample filtering**	**Detected in half-sample filtering**	**Detected in all-sample filtering**	**Predicted ClassyFire Class**	**Description derived from PubChem and HMDB**	**Present in Human Fecal Metabolome Database**	**ReDU Chemical Explorer Associations**
Hypoxanthine	137.046	0.39	0.98	Yes	Yes	No	-	Purine derivative associated with inosine and uric acid in humans	Yes	Found in bacterial cultures (*Clostridium orbiscindens* CC43_001K)
Nicotinamide N-oxide	139.05	0.31	1	Yes	No	No	-	Precursor to nicotinamide -adenine dinucleotide (NAD+) in animals	No (found in blood)	Associated with *Staphylococcus lugdunensis*. Found in human skin, saliva, and fecal samples
3-methyl-2-oxindole (3-Methyloxindole)	148.076	3.22	0.93	Yes	No	No	Indoles and derivatives (55.56%)	Endogenous product of 3-methylindole	No	Found in bacterial cultures (*Gardnerella vaginalis* and *Collinsella* sp. 4_8_47FAA)

Hyocholic acid	158.154	4.78	0.83	Yes	Yes	No	Steroids and steroid derivatives (100%)	Mammalian bile acid	Yes	Largely found in bacterial culture and humans
Gly-Val (Glycylvaline)	175.107	0.36	0.91	Yes	Yes	Yes	-	Glycine and valine dipeptide	Yes	Found in bacterial cultures, associated with *Staphylococcus aureus*. Also associated with inflammatory bowel disease.
3-Hydroxy-4-methoxycinnamic acid (Isoferulic acid)	**177.055**	**4.13**	**0.97**	**Yes**	**Yes**	**No**	**Cinnamic acids and derivatives (9.09%)**	**Endogenous human metabolite; Also potential biomarker for coffee, wheat, sunflowers, etc.**	**Yes**	**Predominantly found in plant, food (such as fruits), and beverage samples. Also found in human caecum and fecal samples.**

2-Butanone, 4-(2,6,6-trimethyl-2-cyclohexen-1-yl)	**177.164**	**3.15**	**0.83**	**Yes**	**No**	**No**	**-**	**Also known as 3α,7α,12α-trihydroxycholestanoic acid (THCA); Intermediate bile acid associated with metabolic disorders**	**No**	**Predominantly found in plant samples (*Cucumis melo*), bacterial culture (*Bacteroides dorei*; *Propionibacterium acnes*), and fungal cultures.**
Paraxanthine	181.072	0.81	0.94	Yes	Yes	No	Imidazopyrimidines (21.74%)	Primary metabolite of caffeine; Found in animals and some bacteria	Yes	Not in Chemical Explorer
trans-Ferulic acid	195.065	3.01	0.91	Yes	Yes	No	Cinnamic acids and derivatives (9.09%)	Abundant in plant cell walls	Yes	Predominantly found in plant, food, and beverage samples. Also found in human fecal and caecum samples.
Loliolide	197.117	3.11	0.94	Yes	Yes	No	Benzofurans (13.79%)	Plant metabolite	No	Found in environmental and plant

										samples (*Cucumis melo*)
3-Hydroxydodecanoic acid	199.169	5.35	0.99	Yes	Yes	No	Lactones (20%)	Medium-chain fatty acid	Yes	Found in environmental samples, plant samples, and some human feces
N-Acetyl-D-mannosamine	204.087	3.27	0.91	Yes	No	No	-	Bacterial metabolite; Precursor to *N*-acetylmannosamine	Yes	Found in a fungal culture (species not specified)
N-acetyl-L-Phenylalanine	208.097	2.84	0.95	Yes	Yes	No	Carboxylic acids and derivatives (10.39%)	Metabolite of phenylalanine; Frequently found in urine of humans with phenylketonuria	Yes	Found in bacterial (*Fusobacterium ulcerans* 12-1B, *Sutterella wadsworthensis* HGA0223, etc.) and some fungal cultures

Thr-Pro (Threonylproline)	217.122	0.46	0.96	Yes	Yes	No	-	Threonine and proline dipeptide	No	Found in bacterial culture, fungal culture, and animal samples (e.g., mice, rats, humans, etc.). Found in human fecal samples. Associated with inflammatory bowel disease samples. Also associated with *Staphylococcus aureus*.
Val-Val (Valylvaline)	217.155	0.45	0.97	Yes	Yes	Yes	Carboxylic acids and derivatives (12.86%)	Valine and valine dipeptide	Yes	Found in fungal cultures and human fecal samples.
Abrine	219.113	0.60	0.88	Yes	Yes	No	-	Associated with *Escherichia coli* metabolism	No	Found in food, beverage, plant, animal, fungal culture, and bacterial culture samples. Also found in human digestive tract, fecal, heart,

										caecum, skin, and blood samples. Found equally in human rural and urban samples. Associated with many different bacterial species (e.g., *Parabacteroides goldsteinii* CC87F, *Prevotella nigrescens* CC14M, *Bifidobacterium longum*, etc.)
Pantothenic acid	**220.118**	**0.56**	**0.81**	**Yes**	**No**	**No**	**-**	**Also called Vitamin B5; essential metabolite for carbohydrate, protein, and fat synthesis**	**Yes**	**Found in animal (human, mouse, rat, etc.) and environmental samples.**

PyroGlu-Pro (Pyroglutamylproline)	227.103	0.44	0.86	Yes	Yes	Yes	-	No information provided	No	Predominantly found in bacterial cultures (*Prevotella nigrescens, Bacteroides stercoris* CC31F, etc.), plant, and human fecal samples
Myristoleic acid	227.201	5.94	0.94	Yes	Yes	Yes	-	Long-chain fatty acid found in all eukaryotes; Potential biomarker for some dairy products and other food (anchovies,	Yes	Predominantly found in environmental samples, such as soil.

								dates, sunflowers, chocolate, etc.)		
Ile-Pro (Isoleucylproline)	229.155	0.7	0.98	Yes	No	No	-	Isoleucine and Proline dipeptide	No (found in blood and sweat)	Predominantly found in bacterial (e.g., *Prevotella denticola; Bacteroides stercoris; Parabacteroides johnsonii*, etc.) and fungal cultures
Val-Ile (Valylisoleucine)	231.171	2.82	0.88	Yes	Yes	Yes	-	Valine and Isoleucine dipeptide	Yes	Found in fungal cultures
cis-9-Hexadecenoic acid (Palmitoleic acid)	**237.221**	**6.49**	**0.97**	**Yes**	**Yes**	**Yes**	**-**	**Unsaturated fatty acid**	**Yes**	**Predominantly found in soil samples**
Gly-Tyr (Glycyltyrosine)	239.102	0.44	0.95	Yes	Yes	No	-	Glycine and tyrosine dipeptide	Yes	Found in bacterial cultures (e.g., *Parvimonas micra*, *Fusobacterium*

										nucleatum, etc.)
Biotin	245.098	2.39	0.87	Yes	No	No	Biotin and derivatives (16.67%)	Also called vitamin H; essential human metabolite	Yes	Found in environmental, animal, and bacterial culture (e.g., *Escherichia coli*, *Staphylococcus aureus*, etc.) samples. Also found in human nasal cavity, skin, and saliva samples.

Leu-Leu (Leucylleucine)	245.186	2.40	0.98	Yes	Yes	Yes	Carboxylic acids and derivatives (12.86%)	Leucine and leucine dipeptide	Yes	Predominantly found in bacterial cultures (*Bacteroides stercoris* CC31F, *Clostridium cadaveris* CC44_001G, etc.) and fungal cultures (not specified). Also found in human fecal samples.
Lenticin	247.145	1.22	0.98	Yes	No	No	-	Found in lentil extracts; Possible lentil biomarker	Yes	Found in food, animal, and bacterial culture samples (e.g., *Staphylococcus aureus*). Also found in human urine, milk, blood, and saliva samples.

Val-Met (Valylmethionine)	249.126	0.56	0.95	Yes	Yes	No	-	Valine and Methionine dipeptide	Yes	Not in Chemical Explorer
Ser-Phe (Serylphenylalanine)	253.118	0.79	0.93	Yes	Yes	Yes	-	Serine and phenylalanine dipeptide	Yes	Found in fungal cultures, bacterial cultures (*Bacteroides stercoris* CC31F, *Clostridium cadaveris* CC88A, etc.) and human fecal samples
Palmitelaidic acid	**255.232**	**5.96**	**0.92**	**Yes**	**No**	**No**	**-**	**Trans fatty acid**	**Yes**	**Found in human colon, upper digestive tract, liver, and fecal samples. Also found in soil samples.**

L-Saccharopine	259.129	0.31	0.78	Yes	Yes	No	-	Involved in lysine degradation	Yes	Found in bacterial culture (*Staphylococcus aureus* JE3) and fungal cultures (not specified)
Phe-Pro (Phenylalanylproline)	263.139	2.57	0.99	Yes	Yes	Yes	-	Phenylalanine and Proline dipeptide	Yes	Predominantly found in bacterial cultures (*Bacteroides dore*i CL03T12C01, *Prevotella histicola*, *Parabacteroides goldsteinii* CC87F, etc.) and some fungal cultures
Conjugated linoleic Acid (10E,12Z)	**263.237**	**7.52**	**0.85**	**Yes**	**Yes**	**No**	**Prenol lipids (5.36%)**	**Variation of conjugated linoleic acid; Found in meat dairy products of ruminants; Dietary supplement**	**No**	**Found in human digestive tract, liver, colon, and feces. Also found in environmental and fungal samples.**

Conjugated linoleic acid (9E,11E)	263.237	6.68	0.92	Yes	Yes	Yes	Prenol lipids (5.36%)	Variation of conjugated linoleic acid; Found in meat dairy products of ruminants; Dietary supplement	No (found in blood)	Found in environmental samples and fungal cultures. Also found in human upper digestive tract, liver, and feces.
Thr-Phe (Threonylphenylalanine)	267.134	0.48	0.94	Yes	Yes	Yes	-	Threonine and Phenylalanine dipeptide	Yes	Found in fungal cultures
Pro-Arg (Prolylarginine)	272.171	0.32	0.72	Yes	No	No	-	Proline and arginine dipeptide	No	Found in a fungal culture (species not specified)
9-OxoOTrE	275.201	4.42	0.8	Yes	Yes	No	-	Long-chain fatty acid	No	Found in fungal cultures and plant samples (*Zea mays* L. and *Cucumis melo*)

Phe-Leu (Phenylalanylleucine)	279.171	2.59	0.99	Yes	Yes	Yes	Carboxylic acids and derivatives (12.86%)	Phenylalanine and leucine dipeptide	No	Found in human upper digestive tract, colon, and fecal samples. Also found in bacterial cultures (*Parabacteroides goldsteinii* CC87F, *Prevotella nigrescens* CC14M, etc.) and fungal cultures (not specified)
Leu-Phe (Leucylphenylalanine)	279.171	3.47	0.98	Yes	Yes	Yes	-	Leucine and phenylalanine dipeptide	Yes	Found in animal, fungal culture, and bacterial culture (e.g., *Bacteroides dorei*, *Staphylococcus aureus*, *Propionibacterium acidifaciens*, etc.) samples. Also found in human intestinal,

										vaginal, and fecal samples
N-Tetracosenoyl-4-sphingenine	282.279	6.09	0.95	Yes	Yes	No	Fatty Acyls (11.54%)	Ceramide associated with cell physiology and some human pathologies	Yes	Not in Chemical Explorer
Octadecanamide	284.295	8.55	0.77	Yes	No	No	-	Metabolite derived from stearic acid, found in plant and animal fats	Yes	Found in environmental and plant samples
Xanthosine	285.083	0.40	0.9	Yes	Yes	No	-	Purine nucleoside	Yes	Found in built environment, animal, and bacterial culture samples.
Arg-Ile (Arginylisoleucine)	288.203	0.36	0.96	Yes	Yes	No	-	Arginine and isoleucine dipeptide	Yes	Found in food and animal samples. Also found in human stomach, saliva, spleen, and fecal samples.

cis-11,14-Eicosadienoic acid	291.268	5.54	0.79	Yes	No	No	Prenol lipids (5.36%)	Omega-6 fatty acid found in human milk	Yes	Found in bacterial culture (*Propionibacterium acnes*) and plant samples
N-Acetylmuramic Acid	294.119	0.38	0.84	Yes	Yes	Yes	Carboxylic acids and derivatives (4.08%)	Component of bacterial cell walls	Yes	Found in animal samples. In humans, found in caecum and fecal samples.
Tyr-Leu (Tyrosylleucine)	**295.165**	**2.71**	**0.97**	**Yes**	**Yes**	**Yes**	**Carboxylic acids and derivatives (31.25%)**	**Tyrosyl and leucine dipeptide**	**Yes**	**Found in fungal and bacterial cultures (*Prevotella nigrescens* CC14M). Also found in human feces**

Phe-Met (Phenylalanylmethionine)	297.126	2.18	0.93	Yes	Yes	No	-	Phenylalanine and methionine dipeptide	Yes	Found in bacterial cultures (*Bacteroides stercoris* CC31F, *Prevotella bivia*, etc.).
Ile-Gly-Ile (Isoleucylglycylisoleucine)	302.205	3.04	0.94	Yes	Yes	No	-	Isoleucine, glycine, and isoleucine tripeptide	No	Found in animal and bacterial culture samples (*Staphylococcus aureus*, *Bacteroides dorei*, *Bacteroides stercoris*, etc.). Also found in human samples collected from different body parts (duodenum, jejunum, urine, ileum, colon, saliva, stomach, fecal, etc.)

Val-Trp (Valyltryptophan)	304.167	3.02	0.97	Yes	Yes	No	-	Valine and tryptophan dipeptide	Yes	Found in fungal culture, food, and animal samples. Also found in human jejunum, stomach, duodenum, vagina, spleen, ileum, and fecal samples
Fructoselysine	309.164	0.31	0.85	Yes	No	No	Carboxylic acids and derivatives (4.08%)	Potential biomarker for milk and milk products	No (found in blood)	Found in animal samples. Small percentage also found in food samples. Found in human jejunum, kidney, ileum, fecal, and blood samples
N-Palmitoylglycine	314.27	7.34	0.86	Yes	No	No	-	Human metabolite with fatty acid group	No	Found in environmental and animal samples.

Ile-Trp (Iseoleucyltryptophan)	318.167	2.27	0.97	Yes	Yes	Yes	-	Isoleucine and tryptophan dipeptide	Yes	Found in bacterial cultures (*Bacteroides stercoris* CC31F, *Selenomonas noxia*, etc.). Small percentage also in animal and fungal cultures. Also found in human intestinal samples
Lithocholic acid	**323.273**	**6.84**	**0.97**	**Yes**	**Yes**	**Yes**	**Prenol lipids (11.11%)**	**Secondary bile acid**	**Yes**	**Found in human feces.**
Leucine enkephalin	336.192	3.22	0.77	Yes	Yes	Yes	-	Enkephalin peptide; Produced in brain	No (found in blood)	Found in bacterial cultures (*Parabacteroides goldsteinii* CC87F, *Parabacteroides merdae*, etc.)

13-Docosenamide, (Z)- (Erucamide)	338.342	9.19	0.83	Yes	Yes	Yes	-	Fatty amide	No	Found in human colon, upper digestive tract, liver, and fecal samples. Also found in bacterial cultures (*Peptostreptococcus sp.* CC14N, *Bifidobacterium longum subsp. longum* 44B, etc.) and fungal cultures
Phe-Trp (Phenylalanyltryptophan)	352.166	3.28	0.96	Yes	Yes	No	-	Phenylalanine and tryptophan dipeptide	Yes	Found in fungal cultures (species not specified)
Ile-Val-Lys (Isoleucylvalyllysine)	359.266	0.60	0.91	Yes	Yes	Yes	-	Isoleucine, valine, and lysine tripeptide	No	Found in animal and bacterial culture samples. Also found in human jejunum, duodenum, ileum, caecum, colon, vaginal, and fecal samples

Cholesterol	**369.352**	**10.50**	**0.97**	**Yes**	**Yes**	**Yes**	**-**	**Animal sterol from body tissues and plasma**	**Yes**	**Found in plant and environmental samples.**
(R)-4-((3R,5S,8R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-7,12-dioxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid	**405.264**	**4.81**	**0.92**	**Yes**	**Yes**	**No**	**Steroids and steroid derivatives (46.15%)**	**NA**	**No**	**Predominantly found in bacterial cultures (*Bacteroides caccae*, *Bacteroides ovatus*,*Clostridium orbiscindens*, etc.). Also found in mouse and human digestive tract and fecal samples.**
Glycoursodeoxycholic acid	**414.301**	**5.01**	**0.9**	**Yes**	**Yes**	**No**	**Steroids and steroid derivatives (36.71%)**	**Secondary bile acid**	**Yes**	**Found in animal samples (rats and humans). Also found in human blood plasma, blood serum, urine, skin, and fecal samples.**

Cholic acid	**426.318**	**5.19**	**0.98**	**Yes**	**Yes**	**Yes**	**Steroids and steroid derivatives (8%)**	**Bile acid produced in liver**	**Yes**	**Found in human GI tract and feces. Associated with urban samples. Also found in bacterial cultures**
6R)-2-(hydroxymethyl)-6-((3R,5R,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)heptanoic acid	431.318	5.25	0.86	Yes	Yes	No	Steroids and steroid derivatives (6.12%)	NA	No	Found in human feces (<18 years old)

Oleanolic acid	439.359	7.62	0.92	Yes	No	No	-	Plant metabolite	Yes	Found in fungal culture, food, plant, environment, and beverage samples. In humans, found in caecum and fecal samples.
Glycocholic acid	**466.32**	**4.69**	**0.96**	**Yes**	**Yes**	**No**	**Steroids and steroid derivatives (13.04%)**	**Secondary bile acid**	**Yes**	**Found in bacterial culture (e.g., *Enterococcus faecium*, *Prevotella oralis*, etc.) and food samples. In humans, found in urine, jejunum, digestive tract, and fecal samples.**

Enoxolone	**471.347**	**6.92**	**0.86**	**Yes**	**No**	**No**	**-**	**Derived from plant metabolite; Commonly used as artificial sweetener**	**No (found in blood)**	**Found in human feces (largely individuals <18 years) and fungal cultures**
Bilirubin	**585.272**	**8.90**	**0.95**	**Yes**	**Yes**	**No**	**-**	**Bile pigment produced during heme breakdown**	**Yes**	**Found in upper digestive tract and feces of humans.**
Urobilin	**591.318**	**4.07**	**0.96**	**Yes**	**Yes**	**Yes**	**-**	**Responsible for yellow coloring of urine**	**Yes**	**Found in urban human colon and fecal samples.**
Stercobilin	**595.349**	**4.05**	**0.94**	**Yes**	**Yes**	**No**	**Tetrapyrroles and derivatives (53.85%)**	**Responsible for brown coloring of feces**	**Not available**	**Found in human fecal samples.**

Table S2.3: Metabolite searches from MASST and ReDU co-analysis with all available human fecal samples.
MASST results contain total dataset matches and common recurring sample type matches, such as mouse, rat, environment, etc. Approximate counts for dataset types, such as plants, are indicated by "~", For ReDU All Human Fecal Sample Search, n=5,556 (our 90 samples + 5,466 ReDU fecal samples). Matched bacterial species names were provided if the bacterial species appeared once in the matches, and otherwise, only genus name was recorded (e.g., Streptomyces means multiple Streptomyces matches occurred for that metabolite). "NA" values represent no results. Matched MASST dataset titles included study name as listed in GNPS, with cited publications (if available).

Compound Name	***m/z***	**RT (min)**	**MASST Results**	**Common MASST Bacterial matches**	**Common MASST dataset matches**	**MASST Link**	**Present in ReDU All Human Fecal Sample Search**
Hypoxanthine	137.046	0.39	**5 dataset matches**; 1 human, 4 mouse	NA	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4a2ea5870d2e4622a5b90e4edfca3f17	Present
Nicotinamide N-oxide	139.05	0.31	**5 dataset matches**; 1 human, 4 mouse	NA	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6cd1ad150059472e84f3f958425880083	Absent
3-methyl-2-oxindole (3-Methyloxindole)	148.076	3.22	No results	NA	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1	Present

						8ca57b672484cb2b4d89b5235afc96d	
Hyocholic acid	158.154	4.78	**12 dataset matches**; 1 human, 1 mouse, 3 environmental samples	*Pseudomonas, Streptomyces albus, Bacillus subtilis, E. coli, Bacteroides, Eubacterium, Akkermansia*	Human Gut Bacteria Biotransformation	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b93deedfe65c46de8b656c93f0e7f50d	Absent
Gly-Val (Glycylvaline)	175.107	0.36	**7 dataset matches**; 1 human, 1 environmental sample	*Pseudomonas, Streptomyces, Staphylococcus aureus*	Zoo mammal fecal metabolomes[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1ce565d3cbe842b1a3f3b82de47606a7	Present
3-Hydroxy-4-methoxycinnamic acid (Isoferulic acid)	177.055	4.13	**18 dataset matches**; 1 human, 5 mouse, 1 environmental sample, 4 plant (sweet orange, tomato), 1 mushroom dataset	"Bacteria"	Human-associated bacteria cultured with bile acids, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=10153589289f45448d9db441d300369a	Absent

2-Butanone, 4-(2,6,6-trimethyl-2-cyclohexen-1-yl)	177.164	3.15	**39 dataset matches**; 6 human, 4 mouse, 5 rat, 4 environmental samples, 6 plant (rosemary, *Eudicotyledons*, thale cress, tomato)	*Marinobacter*, *Amycolatopsis*, *Streptomyces*	Burger ingredients common human diet, Amerindians urbanization gradient, zoo mammal fecal metabolomes[231], anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d762844f0b444356a4958da4c610765f	Present
Paraxanthine	181.072	0.81	**14 dataset matches**; 11 human, 1 mouse, 1 environmental samples	NA	Anti-inflammatory diet in rheumatoid arthritis patients, Nobel twin study, Human rheumatoid arthritis, Metabolomic analysis of Alzheimer's disease	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fc336a940189429f85395db848e77d57	Present

trans-Ferulic acid	195.065	3.01	**52 dataset matches**; 11 human, 20 mouse, 1 rat, 1 environmental sample, plants (melon, corn, tomato, oranges), chemical standards with human samples	*Bacillus subtilis, Pseudomonas, Bacteroides, Bifidobacterium, Akkermansia*	Human Gut Bacteria Biotransformation	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=bf38c3b5b06c440ba397bbc10cf73601	Present
Loliolide	197.117	3.11	**95 dataset matches**; 10 human, 4 mouse, 2 rat, 28 environmental samples, ~10 plants (thale cress, tomato, thapsia, melon, corn)	*Pseudomonas, Bacillus, Cyanobacteria, Streptomyces*	Zoo mammal fecal metabolomes[231], Amerindians urbanization gradient, ONR Human Wright Study, Analysis of Alzheimer's	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=be50e86038794e1b9e0225109de958fa	Present
3-Hydroxydodecanoic acid	199.169	5.35	**21 dataset matches**; 13 human, 3 mouse, 1 environment	*Bacillus subtilis, Burkholderia*	American Gut (with comparison for high and low plant	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fb1b298fcd564e1	Present

			al samples, 1 plant (corn), 1 food fermentation metagenome		consumption), Nobel twin study, Human fecal material standards, Human rheumatoid arthritis	594389d0207087f21	
N-Acetyl-D-mannosamine	204.087	3.27	**83 dataset matches**; 20 human, 11 mouse, 2 rat, 20 environmental samples, ~4 plant (tomato, cotton)	*Streptomyces albus*, *Pseudomonas*, *E. coli*, *C. difficile*, *Buckholderia*, *Actinomycete*	American Gut, Amerindians Urbanization Gradient, Human fecal characterization test, Human rheumatoid arthritis	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fc40869d5519485a808dc7caf0a2b8ae	Present
N-acetyl-L-Phenylalanine	208.097	2.84	**76 dataset matches**; 17 human, 15 mouse, 7 environmental samples, 8 food metagenome, 1 plant (tomato)	*Actinobacteria*, *Streptomyces*, *Rothia isolates*, *Bacillus*, *Akkermansia*, *Bifidobacterium*, *Pseudomonas*, *Cutibacterium acnes*, *Staphylococcus aureus*	American Gut (also with high and low plant consumption comparison), Amazon urbanization housing analysis, Nobel twin study, Human Gut Bacteria Biotransformation	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7279069e65014373ad790ecb80156672	Present

Thr-Pro (Threonylproline)	217.122	0.46	**15 dataset matches**; 5 human, 5 mouse, 1 rat, 1 environmental samples, 2 plants (tomato)	"Bacteria"	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=607f9ac84efe43a9a26a642389869975	Present
Val-Val (Valylvaline)	217.155	0.45	**35 dataset matches**; 6 human, 7 mouse, 4 rat, 2 environmental samples, 2 food metagenome, ~3 plant (thale cress, potato, wild oat)	*Staphylococcus aureus, Bacteroides, Akkermansia, Bifidobacterium*	Human gut bacteria biotransformation, Zoo mammal fecal metabolomes[231], Anti-inflammatory diet in rheumatoid arthritis patients, Human fecal material standards, Burger ingredients, ONR feces and plasma	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a0be1a5518ef4594bd188eed2bd4e6d2	Present

Abrine	219.113	0.61	**3 dataset matches**; 3 human	NA	American Gut (also UK), Malawi legume study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7a23704e8e7c40a7a18692083e5b00c2	Present
Pantothenic acid	220.118	0.56	**142 dataset matches**; 33 human, 33 mouse, 4 rat, 5 environmental samples, ~11 plant (potato, peppers, melon, corn, sugarcane, tomato, thale cress, thapsia, *Nicotiana benthamiana*)	*Pseudomonas*, *E. coli*, *Streptomyces*, *Bacillus*, *Staphylococcus aureus*	American gut (also UK; also high vs low plant consumption comparison), Nobel twin study, Zoo mammal fecal samples[231], Human rheumatoid arthritis	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=bce374bc20ef4f69ae9348a776d48e8f	Present

PyroGlu-Pro (Pyroglutamylproline)	227.103	0.44	**44 dataset matches**; 20 human, 5 mouse, 7 food metagenome, 2 environmental samples, ~3 plant (cotton, *Nicotiana benthamiana*, corn)	*Amycolatopsis, Bacillus subtilis, Pseudomonas, Streptomyces albus*	American Gut, ONR human diet, Human-associated bacteria cultured with bile acids	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=893f768b31de46f5903ad003dfdb2152	Present
Myristoleic acid	227.201	5.94	**35 dataset matches**; 17 human, 6 mouse, 1 rat, 3 environmental samples, ~2 plant (cotton, tomato)	*Streptomyces chartreusis*	Burger ingredients, American gut (with high vs low plant consumption comparison), Amerindians urbanization gradient, ONR Primary Wright, Human fecal material standards, Alzheimer's analysis	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=09d5bd98a46145e982b986a34e4591ef	Present

Ile-Pro (Isoleucylproline)	229.155	0.76	**168 dataset matches**; 44 human, 32 mouse, 5 rat, 12 environmental samples, 7 food metagenome, ~9 plant (tomato, corn, cotton, potato)	*Pseudomonas*, *Streptomyces*, *Bacillus*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Staphylococcus aureus*, *Lactobacillus*, *Amycolatopsis*, *Paenibacillus*, *E. coli*, *Eggerthella lenta*, *Serratia plymuthica*	American Gut (also UK and with plant consumption comparison), Human gut bacteria biotransformation, Human-associated bacteria cultured with bile acids, Zoo mammal fecal samples[231], Nobel twin study, Burger ingredients, Amerindians urbanization gradient, Human skin microbe isolates[232]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=63a8a19454df4473b3bcbab0a4023ce9	Present
Val-Ile (Valylisoleucine)	231.171	2.82	**80 dataset matches**; 18 human, 11 mouse, 4 rat, 3 environmental samples, 6 food metagenome	*Staphylococcus*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Pseudomonas*, *Actinomycete*, *Streptomyces*, *Alteromonas*	Human gut bacteria biotransformation, American gut (also UK), ONR Primary Wright, Burger ingredients, Human fecal	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c96f435a91334e1fb38bcd1d0c111a11	Present

			, ~6 plant (thale cress, *Nicotiana benthamiana*, tomato, cotton)		material standards		
cis-9-Hexadecenoic acid (Palmitoleic acid)	237.221	6.49	**73 dataset matches**; 33 human, 14 mouse, 4 rat, 5 environmental samples (also built environment), ~6 plant (tomato, corn, sugarcane)	*Bacillus subtilis*, *Bacteroides*, *Bifidobacterium*, *Akkermansia*, *Lactobacillus*, *Cyanobacteria*, *Marinobacter*, *Prochlorococcus marinus*, *Rothia*	Human gut bacteria biotransformation, American Gut (subset for plant consumption comparison), ONR primary Wright, burger ingredients, Amerindians urbanization gradient, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d3eed7e162f24c3ab3b2c1c3e2e98785	Present
Gly-Tyr (Glycyltyrosine)	239.102	0.44	**34 dataset matches**; 13 human, 10 mouse, 1 rat, 3 environmental samples, ~2 plant (tomato, potato)	*Staphylococcus aureus*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*	Human gut bacteria biotransformation, Human-associated bacteria cultured with bile acids, anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ecdd1a94b14a4822a4466ca3c86b4bab	Present

Biotin	245.098	2.39	No results	NA	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7e05ee9c085e492a871eb3bea1d24f50	Present
Leu-Leu (Leucylleucine)	245.186	2.40	**190 dataset matches**; 60 human, 30 mouse, 5 rat, 16 environment (also built environment), 13 food metagenome, ~14 plant (tomato, potato, melon, corn, malus, cotton, *Nicotiana benthamiana*, oranges)	*E. coli*, *Staphylococcus aureus*, *Pseudomonas*, *Streptomyces*, *Nocardia*, *Bacillus*, *Burkholderia*, *Eggerthella lenta*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Lactobacillus*, *Roseburia*, *Paenibacillus*, *Alteromonas*, *Amycolaptosis*, *Xanthomonas citri*, *Actinomycete*	American gut (with plant consumption comparison), Human gut bacteria biotransformation, Burger ingredients, Human-associated bacteria cultured with bile acids, ONR human diet, ONR primary wright, Human skin microbiome isolates[232], Amazon	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=5362cc2f00994eeba80d4bc3db7954e0	Present

					urbanization housing analysis, Zoo mammal fecal samples[231], Ancient DNA teeth, Anti-inflammatory diet of rheumatoid arthritis patients, Amazon skin and environmental samples		
Lenticin	247.145	1.22	**42 dataset matches**; 35 human, 3 environmental samples, 1 plant (tomato)	NA	ONR Human study, Human rheumatoid arthritis, American gut (also UK), Zoo mammal fecal samples[231], Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=da61e8f7b5e744ac99a829e23a067c71	Present

Val-Met (Valylmethionine)	249.126	0.56	**27 dataset matches**; 4 human, 6 mouse, 5 rat, 3 food metagenome, ~2 plant (potato, tomato)	*Staphylococcus aureus, Bacteroides, Akkermansia, Bifidobacterium*	Human gut bacteria biotransformation, Burger ingredients, ONR human diet, Zoo mammal fecal samples[231], Anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3772722082864e3db8ffa8be8d87607c	Present
Ser-Phe (Serylphenylalanine)	253.118	0.79	**64 dataset matches**; 18 human, 13 mouse, 2 rat, 9 environmental samples, ~6 plant (cotton, thale cress, potato, corn, tomato)	*Pseudomonas, Streptomyces, E. coli, Amycolaptosis, Bacillus*	Human gut bacteria biotransformation, American gut (with plant consumption comparison), Zoo mammal fecal samples[231], Burger ingredients, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4cd767ae2c9d4f2a815c2120511c9112	Present
Palmitelaidic acid	255.232	5.96	**31 dataset matches**; 12 human, 9 mouse, 1 rat, 3 environment	*Bacillus subtilis*	Amerindians urbanization gradient	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8dbc4bee89914f249f91ab103ac05d26	Present

			(also built environment), ~1 plant (corn)				
L-Saccharopine	259.129	0.31	**15 dataset matches**; 5 human, 2 mouse, ~1 plant (tomato)	*Streptomyces clavuligerus*, *Streptomyces albus*	American and UK gut	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e4dd2264edc145dd9fd0525d9e8fbc4a	Present
Phe-Pro (Phenylalanylproline)	263.139	2.57	**196 dataset matches**; 62 human, 29 mouse, 5 rat, 18 environment (including built environment), 8 food metagenome, ~4 plant (cotton, sugarcane, tomato, corn)	*Staphylococcus aureus*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Bifidobacterium*, *E. coli*, *Pseudomonas*, *Bacillus*, *Burkholderia*, *Amycolaptosis*, *Leptospira*, *Eggerthella lenta*, *Actinomycete*, *Paenibacillus*, *Streptomyces*, *Acinetobacter pittii*, *Pantoea conspicua*, *Klebsiella*, *Enterobacter*,	Human gut bacteria biotransformation, American gut (with plant consumption comparison), Human-associated bacteria cultured with bile acids, Burger ingredients, Human skin microbiome isolates[232], Human rheumatoid arthritis, Zoo mammal fecal samples[231], Nobel twin	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ae38f3f9e8864fb59edb943a8e514bae	Present

				Rothia, *Veillonella atypica*	study, Amazon skin and environmental samples		
Conjugated linoleic Acid (10E,12Z)	263.237	7.52	**132 dataset matches**; 43 human, 33 mouse, 6 rat, 10 environmental samples (including built environment), 14 food metagenome, ~6 plant (potato, corn, thale cress), ~5 animal (baboon, cheetah)	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Bacillus*, *Cutibacterium acnes*	Human gut bacteria biotransformation, American gut (with plant consumption comparison), Urbanization gradient foods, ONR wright human study, ONR primary wright, ONR human diet, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=83c13baa8dc74e64b18653edd89e8567	Absent
Conjugated linoleic acid (9E,11E)	263.237	6.68	**144 dataset matches**; 48 human, 36 mouse, 6 rat, 11 environment (including built environment), 12 food	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Methylobacter tundripaludum*, *Clostridium*, *Bacillus subtilis*	ONR Primary Wright Human, Human Gut Bacteria Biotransformation, ONR human diet, American Gut (with plant consumption comparison),	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6389f1e6cf4f451893599372f2b682b5	Present

			metagenome, ~4 animal (baboon, rhino, cheetah), ~7 plant (corn, potato, thale cress, oranges)		Amerindians urbanization gradient, Amazon urbanization housing analysis, Nobel twin study, Urbanization gradient foods		
Thr-Phe (Threonylphenylalanine)	267.134	0.48	**44 dataset matches**; 9 human, 10 mouse, 5 rat, 3 environment, ~4 plant (tomato, thale cress, potato), 3 animal (cheetah)	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Staphylococcus aureus*, *Streptomyces*	Human Gut Bacteria Biotransformation, Amerindians urbanization gradient, Human rheumatoid arthritis, Burger ingredients, ONR wright human study, Zoo mammal fecal samples[231], Anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=34e92fc0e92349189ec8234819dac5d8	Present

Pro-Arg (Prolylarginine)	272.171	0.32	**7 dataset matches**; 4 human, 3 mouse	NA	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4dd48325058045aa862fbb31c1a8230d	Present
9-OxoOTrE	275.201	4.42	**58 dataset matches**; 16 human, 8 mouse, 4 rat, 3 environment, 1 food metagenome, ~13 plant (corn, cotton, tomato, *Malpighiaceae* plants, melon, thale cress, potato, orange), ~1 other animal (rhino)	*Bacillus*, *Pseudomonas*, *Streptomyces scabiei*	Burger ingredients, ONR primary wright human study, Human rheumatoid arthritis, Zoo mammal fecal samples[231], Anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9c0b6ad5c2f24f02b49c5d834598b39d	Present

Phe-Leu (Phenylalanylleucine)	279.171	2.59	**222 dataset matches**; 77 human, 37 mouse, 7 rat, 28 environment, 11 food metagenome, ~6 other animal (cheetah, rhino, cat), ~10 plant (cotton, *Malpighiaceae* plants, potato, tomato, thale cress, corn)	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *E. coli*, *Paenibacillus*, *Eggerthella lenta*, *Staphylococcus aureus*, *Pseudomonas*, *Nocardia*, *Streptomyces*, *Veillonella atypica*, *Achromobacter xylosoxidan*, *Cutibacterium acnes*, *Amycolaptosis*	Human gut bacteria biotransformation, American gut (with plant comparison and UK gut), Human-associated bacteria cultured with bile acids, Human skin microbiome isolates[232], Burger ingredients, Urbanization gradient foods, Amerindians urbanization gradient, Amazon urbanization housing analysis, Zoo mammal fecal samples[231], Ancient DNA teeth	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0506cb53c28b4aefa42281c3da9c0881	Present

Leu-Phe (Leucylphenylalanine)	279.171	3.47	**147 dataset matches**; 41 human, 26 mouse, 5 rat, 21 environment, 5 food metagenome, ~3 other animal (baboon, cheetah), ~9 plant (peppers, *Nicotiana benthamiana*, corn, potato, tomato, thale cress, cotton)	*Staphylococcus aureus*, *Eggerthella lenta*, *Akkermansia*, *Bacteroides*, *Bifidobacterium*, *Roseburia*, *Clostridium*, *Bacillus*, *Pseudomonas*, *Amycolaptosis*, *E. coli*, *Nocardia*, *Streptomyces*	Human gut bacteria biotransformation, ONR primary wright, Human-associated bacteria cultured with bile acids, Burger ingredients, Human rheumatoid arthritis, Amerindians urbanization gradient, American gut (with plant consumption comparison), Zoo mammal fecal samples[231], Ancient DNA teeth, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7781fd3196b142e0975dbafc0f737392	Present

N-Tetracosenoyl-4-sphingenine	282.279	6.09	**29 dataset matches**; 7 human, 11 mouse, 4 rat, 2 environment,	"Bacteria"	Burger ingredients, Human rheumatoid arthritis, Amerindians urbanization gradient, Anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ca9486b15bc64efea62c56a21ff74057	Present
Octadecanamide	284.295	8.55	**5 dataset matches**; 3 human, 2 mouse	NA	ONR human plasma	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e11d31d2e9294d8ea30cf1712be3184a	Present

Xanthosine	285.083	0.40	**41 dataset matches**; 5 human, 21 mouse, 3 rat, 1 environment, 3 food metagenome, ~1 plant (*Genipapo*), ~1 other animal (rhino)	*Streptomyces, Pseudomonas fluorescens, Photobacterium galatheae, Paenibacillus peoriae, Xenorhabdus hominickii*	Burger ingredients, Zoo mammal fecal samples[231], Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f40c0769edcc4b89b9e2f5507e2a976a	Present
Arg-Ile (Arginylisoleucine)	288.203	0.36	**53 dataset matches**; 9 human, 9 mouse, 5 rat, 5 environment, 6 food metagenome, ~3 plant (cotton, potato, tomato), ~1 other animal (rhino)	*Bacteroides, Akkermansia, Bifidobacterium, Roseburia, Clostridium, Paenibacillus, Amycolaptosis, Pseudomonas fluorescens, Photobacterium galatheae, Xenorhabdus hominickii*	Human gut bacteria biotransformation, Burger ingredients, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ac1635b5525e4c48a6ba690f9bd314b1	Absent

cis-11,14-Eicosadienoic acid	291.268	5.54	**21 dataset matches**; 7 human, 2 mouse, 1 rat, 2 environment, 1 food metagenome, ~2 plant (Malpighiaceae plants, Myoporeae plants	NA	American Gut (with plant consumption comparison), Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c7075cb219be444493020b722375a145	Absent
N-Acetylmuramic Acid	294.119	0.38	**17 dataset matches**; 1 human, 6 mouse, 1 rat, 2 environment, 1 plant (tomato)	*Streptomyces scabiei*, *Streptomyces*, *Streptomyces clavuligerus*	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6c495d96e0bb4ceeb8c8821fd7a8f57b	Present

Tyr-Leu (Tyrosylleucine)	295.165	2.71	**153 dataset matches**; 48 human, 18 mouse, 6 rat, 16 environment, 12 food metagenome, ~3 other animal (cheetah, dog, rhino), ~6 plant (thale cress, melon, potato, corn)	*E. coli*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Clostridium*, *Pseudomonas*, *Nocardia*, *Streptomyces*, *Bacillus*, *Amycolaptosis*, *Paenibacillus*, *Photobacterium galatheae*, *Xenorhabdus hominickii*	Human Gut bacteria biotransformation, ONR human diet, Human skin microbiome isolates[232], Burger ingredients, Human-associated bacteria cultured with bile acids, ONR primary wright, Amerindians urbanization gradient, Zoo mammal fecal samples[231], Nobel twin study, Anti-inflammatory diet in rheumatoid arthritis patients	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b55595173aa54b429e8056f941be4d3a	Present

Phe-Met (Phenylalanylmethionine)	297.126	2.18	**31 dataset matches**; 5 human, 7 mouse, 3 rat, 4 environment, 4 food metagenome, ~2 plant (tomato, corn)	*Bacillus*	ONR Wright human study, Amerindians urbanization gradient, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f691a68d2503426aa15f80cdd92f7d90	Absent
Ile-Gly-Ile (Isoleucylglycylisoleucine)	302.205	3.04	**99 dataset matches**; 16 human, 29 mouse, 5 rat, 14 environment, 7 food metagenome, ~3 other animal (cheetah, cat, dog), ~5 plant (corn, tomato, potato)	*Eubacteria*, *Bacillus*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Clostridium*, *Pseudomonas*, *Amycolaptosis*, *Streptomyces albus*, *Paenibacillus*	Burger ingredients, Human gut bacteria biotransformation, Human-associated bacteria cultured with bile acids, ONR human diet, Amerindians urbanization gradient, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0b17a25c3d6445d5b2c4a8e335adba23	Present

Val-Trp (Valyltryptophan)	304.167	3.02	**28 dataset matches**; 2 human, 6 mouse, 2 rat, 2 food metagenome ,~3 plant (tomato, cotton, potato), 1 other animal (rhino)	*Streptomyces clavuligerus*, *Streptomyces scabiei*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Clostridium*	Human gut bacteria biotransformation, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2352760f6db441f9bb0880f3ee98c26	Present
Fructoselysine	309.164	0.31	**44 dataset matches**; 12 human, 13 mouse, 4 rat, 1 environment , 3 food metagenome	*Pseudomonas*, *Nocardia*, *Streptomyces*, *Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Clostridium*, *Amycolaptosis*	Human gut bacteria biotransformation, American gut (also UK)	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d3840c7b8fd0438ab2518ebbabbf5b62	Present
N-Palmitoylglycine	314.27	7.34	**8 dataset matches**; 1 human, 2 mouse, ~1 plant (*Euphorbiaceae*)	*Pseudomonas*, *Bacillus*	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=998632092bb7495692e01f7ba1be0248	Present

Ile-Trp (Iseoleucyltryptophan)	318.167	2.27	**59 dataset matches**; 7 human, 12 mouse, 4 rat, 5 environment, 2 food metagenome, ~5 plant (cotton, wild oat, corn, tomato, peppers), ~2 other animal (cheetah)	*Staphylococcus aureus, Bacillus, Streptomyces, Pseudomonas, Amycolaptosis, Bacteroides, Akkermansia, Bifidobacterium, E. coli, Eggerthella lenta*	Burger ingredients, Human gut bacteria biotransformation, Human-associated bacteria cultured with bile acids, ONR Wright human study, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=cf87c54ff6f14611846d4b9bb63f3b3b	Absent
Lithocholic acid	323.273	6.84	**17 dataset matches**; 8 human, 2 mouse, 3 rat, 1 environment, 1 plant (tomato)	NA	American Gut (also UK), ONR Primary Wright, Human rheumatoid arthritis, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2bd41fcad2ef45be8edc72db760066f3	Present

Leucine enkephalin	336.192	3.22	**51 dataset matches**; 8 human, 10 mouse, 2 rat, 3 environment, 5 food metagenome, ~4 plant (corn, tomato), ~2 other animal (cheetah)	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Eggerthella lenta*, *Roseburia*, *lactobacillus*, *Eubacterium*, *Bacillus*, *Streptomyces*, *Pseudomonas*, *Amycolaptosis*, *Paenibacillus*	Human gut bacteria biotransformation, Burger ingredients, ONR human diet, Human skin microbiome isolates[232], IBD severity human samples	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e0bbeeff3ae04f2191b9bedf2ba39a22	Absent

13-Docosenamide, (Z)- (Erucamide)	338.342	9.19	**228 dataset matches**; 52 human, 31 mouse, 6 rat, 24 environment, 5 food metagenome, ~10 other animal (baboon, cheetah, rhino, sheep), ~16 plant (*Mitragyna speciosa*, *Myoporeae* plants, *Angelica keiskei*, *Malpighiaceae* plants, *Thapsia garganica*, oranges, melon, tomato, wild oat, thale cress, *Psychotria insularum*, seaweed)	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Clostridium*, *Methylobacter tundripaludum*, *E. coli*, *Pseudomonas*, *Bacillus*, *Cyanobacteria*, *Colwellia chukchiensis*, *Maribacter ulvicola*, *Algoriphagus machipongonensis*, *Zobellia uliginosa*, *Saccharomonospora viridis*, *Burkholderia*, *Streptomyces*, *Moorena bouillonii*, *Glutamicibacter arilaitensis*, *Staphylococcus aureus*	Human gut bacteria biotransformation, Burger ingredients, ONR primary wright, Human-associated bacteria cultured with bile acids, Human skin microbiome isolates[232], ONR human fecal, American Gut (with plant consumption comparison), Amerindians urbanization gradient, ONR primary bile acids, Ancient DNA teeth	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=50b42b413dbd4aa98745d64374ae8f6d	Present

Phe-Trp (Phenylalanyltryptophan)	352.166	3.28	**43 dataset matches**; 19 human, 11 mouse, 1 rat, 1 environment, ~2 other animal (baboon, cheetah), ~2 plant (tomato)	*Bacillus subtilis, Paenibacillus*	ONR Wright Human study, Burger ingredients, ONR primary wright, Zoo mammal fecal samples[231], Analysis of Alzheimer's, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ab7256224a5240c38e829d2455ec4738	Absent
Ile-Val-Lys (Isoleucylvalyllysine)	359.266	0.60	**13 dataset matches**; 1 human, 5 mouse, 3 rat,	*Pseudomonas, bacillus, Bacillus subtilis*	NA	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f1259278f31d49e7965135dc4fd92f5f	Present

Cholesterol	369.352	10.50	**204 dataset matches**; 102 human, 49 mouse, 6 rat, 5 environment, 17 food metagenome, ~5 plant (seaweed, tomato, sugarcane, thale cress, tomato), ~7 other animal (baboon, dog, sheep)	*Bacteroides*, *Bifidobacterium*, *Akkermansia*, *Cutibacterium acnes*, *Streptomyces coelicolor*, *Pseudomonas aeruginosa*	Amerindians urbanization gradient, Nobel twin study, American Gut (with plant consumption comparison and UK), Human gut bacteria biotransformation, urbanization gradient foods, ONR Wright human study, ONR human diet, ONR human fecal, Molecular analysis of human brain tissues	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4d622370a5fa4ad5bb1eab577469d246	Present

(R)-4-((3R,5S,8R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-7,12-dioxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid	405.264	4.81	**51 dataset matches**; 22 human, 12 mouse, 1 rat, 1 environment, 5 food metagenome, ~3 other animal (cheetah)	*Streptomyces, Pseudonocardia, Gordonia, Kitasatospora*	Human-associated bacteria cultured with bile acids, IBD severity, American Gut (with plant consumption comparison), Zoo mammal fecal samples[231], Nobel twin study, IBD severity	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8cfb07ca20de46af9157341b39d25de5	Present
Glycoursodeoxycholic acid	414.301	5.01	**47 dataset matches**; 26 human, 2 mouse, 5 rat, 3 environment, ~2 other animal (baboon), ~1 plant (tomato)	*Bacteroides, Akkermansia, Bifidobacterium, Alteromonas, Pseudomonas fragi, E. coli, Streptomyces*	Human rheumatoid arthritis, ONR primary wright, Human gut bacteria biotransformation, Nobel twin study, Amerindians urbanization gradient, Analysis of Alzheimer's	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=43e8d71f3ad84cc0b7a7967e821a45a9	Present

Cholic acid	426.318	5.19	**187 dataset matches**; 51 human, 54 mouse, 10 rat, 15 environment, 13 food metageneome, ~2 plant (Ardisia crenata, thale cress), ~6 other animal (baboon, dog, cheetah, sheep)	*Bacteroides*, *Akkermansia*, *Bifidobacterium*, *Roseburia*, *Eggerthella*, *Clostridium*, *Pseudomonas*, *Stenotrophomonas*, *Achromobacter*, *Salinispora*, *Photobacterium galatheae*, *Paenibacillus peoriae*, *Xenorhabdus hominickii*, *Streptomyces*, *Salinispora*, *Nocardiopsis*, *Brevibacterium*, *Bacillus*, *Alteromonas*	ONR Wright Human study, ONR Primary Wright, Burger ingredients, Human gut bacteria biotransformation, American Gut (also UK), Zoo mammal fecal samples[231], Nobel twin study, IBD severity	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=496f2c33aaed4da0a9096623376f7afe	Present
6R)-2-(hydroxymethyl)-6-((3R,5R,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-	431.318	5.25	**6 dataset matches**; 1 human, 2 mouse, 2 rat, 1 other animal (cheetah)	NA	Amerindians Urbanization gradient	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9d7b3242ba97467f88cc7e2849a73c9a	Absent

cyclopenta[a]phenanthren-17-yl)heptanoic acid							
Oleanolic acid	439.359	7.62	**106 dataset matches**; 22 human, 11 mouse, 3 rat, 5 environment, 16 food metagenome, ~5 other animal (cheetah, ant), ~11 plant (rosemary, *Fraxinus excelsior*, oranges, corn, tomato, *Parietaria judaica*)	*Streptomyces*, *Burkholderia*, *Bacillus*, *Eggerthella lenta*, *Cyanonacteria*, *Staphylococcus*, *Acinetobacter pittii*, *Pantoea conspicua*, *Klebsiella*, *Enterobacter*	ONR Wright Human study, American Gut, Amazon urbanization housing analysis, Urbanization gradient foods, Zoo mammal fecal samples[231]	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=800acf7db452461c9d0a890f990ecd83	Present

Glycocholic acid	446.32	4.69	**204 dataset matches**; 85 human, 19 mouse, 2 rat, 2 environment, 18 food metagenome, ~9 plant (corn, tomato, *Ardisia crenata*), ~7 other animal (rhino, baboon, cheetah)	*Cyanobacteria, Pseudomonas, Streptomyces, E. coli, Alteromonas, Bacillus subtilis, Pseudonocardia, Gordonia, Kitasatospora, Actinomadura, Nocardia, Actinomycetes,Pseudoalteromonas, Roseovarius, Staphylococcus aureus, Bacteroides, Akkermansia, Bifidobacterium, Roseburia, Eggerthella, Clostridium, Paenibacillus, Cutibacterium acnes*	Human-associated bacteria cultured with bile acids, ONR Wright human study, IBD human samples, Burger ingredients, Human rheumatoid arthritis, Human gut bacteria biotransformation, Amerindians urbanization gradient, Urbanization gradient foods, ONR human plasma, ONR primary bile acids, Nobel twin study, *Salinispora*, American Gut (with plant consumption comparison)	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8d5d343a95964faba3108020318837ff	Absent

Enoxolone	471.347	6.92	**32 dataset matches**; 12 human, 2 mouse, 2 environment, 4 food metagenome, ~2 other animal (cheetah), ~3 plant (melon)	*Bacillus subtilis, E. coli*	American Gut (also UK), Ancient DNA teeth	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=bdc20710970b4c47a99c178a1c2606e9	Absent
Bilirubin	585.272	8.90	**123 dataset matches**; 65 human, 33 mouse, 7 rat, 1 environment, 1 food metagenome, ~5 other animal (cheetah, baboon, dog), ~2 plant (tomato, thale cress)	*Streptomyces albus, Kitasatospora cystarginea*	ONR human plasma, IBD human fecal samples, ONR Primary Wright, Human rheumatoid arthritis, American gut (with plant consumption comparison), Amerindians urbanization gradient, Amazon urbanization housing analysis, IBD severity	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2405d7eb22f64cceba64d6b43bd2f0aa	Present

Urobilin	591.318	4.07	**116 dataset matches**; 62 human, 34 mouse, 6 rat, 6 environment, ~4 other animal (cheetah, cats, rhino), ~1 plant (tomato)	NA	ONR Primary Wright, ONR Human Study, IBD severity, Human rheumatoid arthritis, American Gut (with plant consumption comparison), Zoo mammal fecal samples[231], IBD severity, Nobel twin study	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=da0037fa7f4c4f72860d39a4d7184402	Present
Stercobilin	595.349	4.05	**94 dataset matches**; 57 human, 16 mouse, 6 rat, 6 environment, 2 food metagenome, ~4 other animal (rhino), ~2 plant (*Genipapo*, tomato)	NA	ONR human fecal, ONR Primary Wright, IBD severity, American gut (also UK), Amazon urbanization housing analysis, Amazon urbanization skin and environmental samples, Zoo mammal fecal samples[231],	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e314b28d4ff343ccbef124eb09ffb016	Present

					Nobel twin study, Anti-inflammatory diet of rheumatoid arthritis patients, Analysis of Alzheimer's disease		

Table S2.4. MZmine parameters for feature-based molecular networking.
**Only done for gap-filled analyses*

MZmine v 2.33

Method	Parameter	Value
Mass Detection	MS1	1.80E+05
	MS2	1.00E+03
Chromatogram Builder	Min time span (min)	0.01
	Min height	5.40E+05
	m/z tolerance (ppm)	1.00E+01
	Filter	MS1
Chromatogram Deconvolution - Local Minimum Search	Chromatographic Threshold	5%
	Search minimum in RT range	0.10 min
	Min relative height	15%
	Min absolute height	3.00E+05
	Min ratio of peak top/edge	2
	Peak duration range	0.01-1.5min
	m/z range for MS2 scan pairing	0.01
	RT range for MS2 scan pairing	0.1min
Isotopic Peaks Grouper	*m/z* tolerance (ppm)	1.00E-02
	Retention time tolerance (min)	0.05
	Max charge	3
	Representative isotope	Lowest *m/z*
	Monotonic shape	Yes
Join Aligner	*m/z* tolerance (ppm)	10
	Weight for *m/z*	5
	Weight for RT	1
	Retention time tolerance (min)	0.5
Peaks List Row Filter	Retention time (min)	0.2-12.51
	Min peaks per row	empty (leave unchecked)
	Keep only peaks with MS2 scan (GNPS)	yes
Gap Filling Peak Finder*	Intensity tolerance (%)	75
	m/z tolerance (ppm)	1.00E+01
	m/z tolerance (*m/z*)	1.00E-06
	Retention time tolerance (min)	0.3
	RT correction	yes

Supplementary Material B

Author List and Affiliations

Yi Sun[1,2#], Jacob J. Haffner[3,4#], Daniela Betancurt[5], Evan R. Hutchison[1], Eugenio I. Vivas[1], Robert L. Kerby[1], Nsoua H. Diaba[6], Camil Gosmanov[6], Alejandro Reyes Muñoz[4], Laura-Isobel McCall[4,6,7*], Federico E. Rey[1,*]

[1]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, 53706, USA

[2]Tianjin Key Laboratory of Food Science and Health, School of Medicine, Nankai University, Tianjin, 300071, China.

[3]Department of Anthropology, University of Oklahoma, Norman, OK 73019, USA.

[4]Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, OK 73019, USA.

[5]Departamento de Ciencias Biológicas, Universidad de los Andes, Bogota, Colombia

[6]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK 73019, USA.

[7]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, USA.

# Authors contributed equally

* Corresponding author(s)

Author Contributions

F.R. and L.-I.M. conceived and designed study. F.R. and Y.S. collected mouse samples, with input from D.B., E.R.H., E.I.V., and R.L.K. J.J.H. conducted metabolite extraction and LC-MS/MS analysis with direction from L.-I.M. J.J.H. and L.-I.M. analyzed untargeted LC-MS/MS data with assistance from N.H.D. and C.G.

Acknowledgements

The authors thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing sequencing and support services.

Supplementary Figures 3.1-3.4

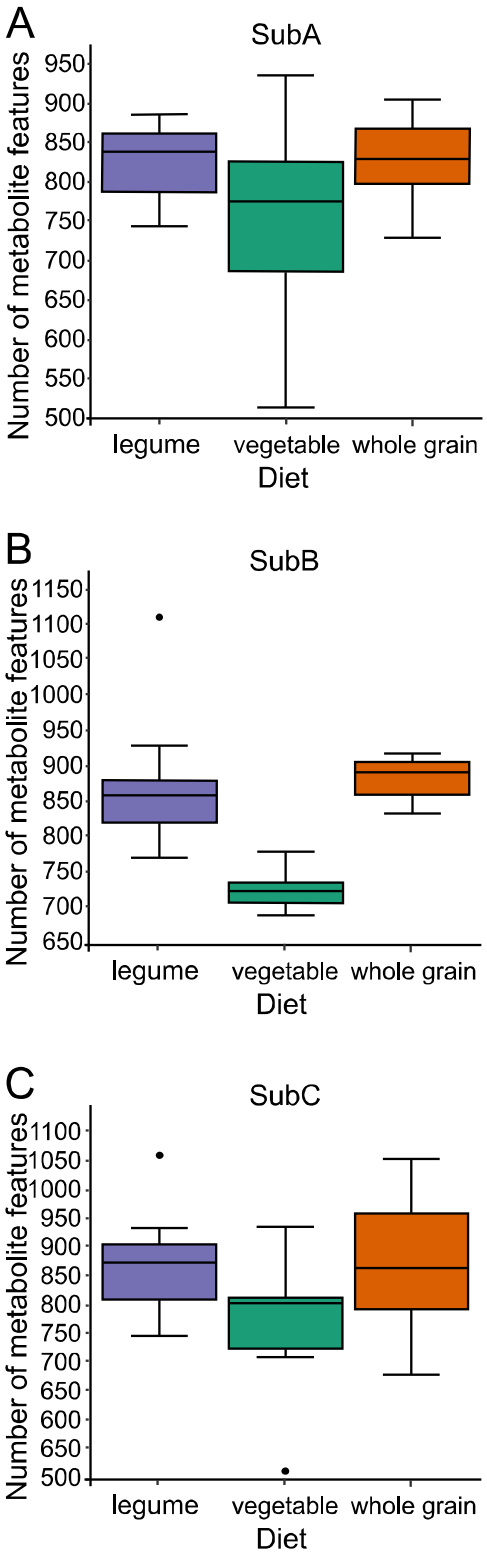


Figure S3.1: Differential impact of diet on alpha diversity of microbiota-modulated metabolites in cecum.
Effects of diet on microbiome communities by (A) SubA, (B) SubB, and (C) SubC.

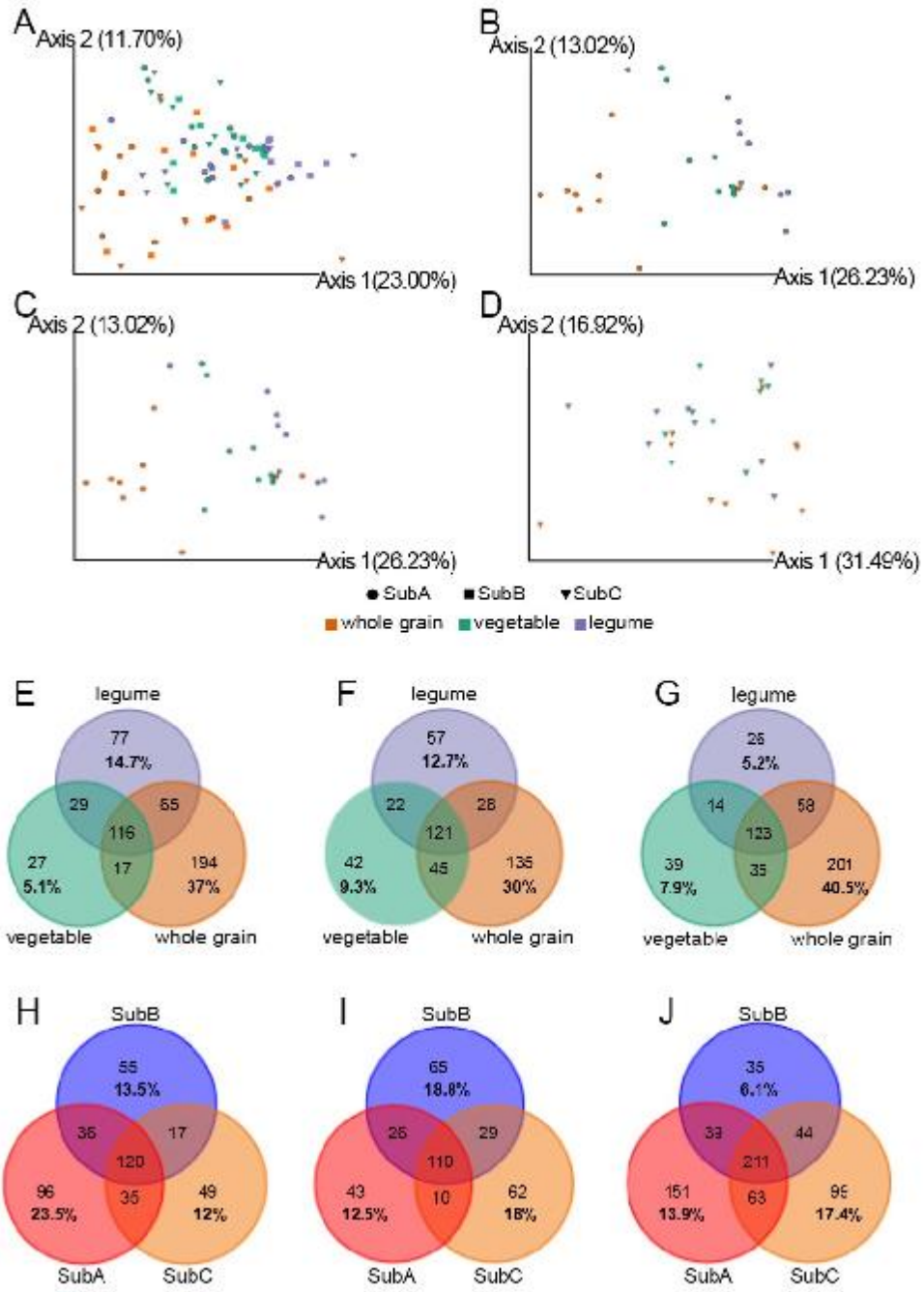


Figure S3.2: Gut community and diet modulate serum metabolites.
(A) PCoA of Bray-Curtis dissimilarities from serum microbial metabolites for all three donors. (B-D) PCoA of Bray-Curtis dissimilarities from serum microbial metabolites for each donor donors. SubA (B), SubB (C) and SubC (D). (E-G) Numbers of elevated/depleted metabolites presented in different diets in each donor. (H-J) Numbers of elevated/depleted metabolites presented in different donors in each diet.

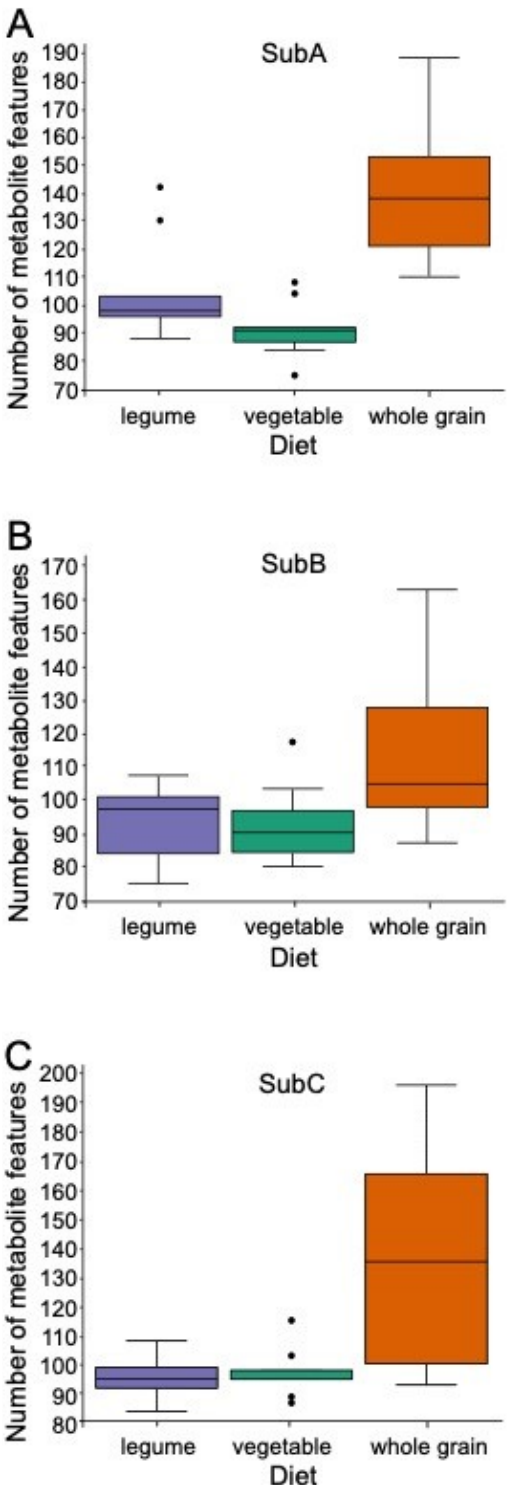


Figure S3.3: Differential impact of diet on alpha diversity of microbiota-modulated metabolites in the serum.

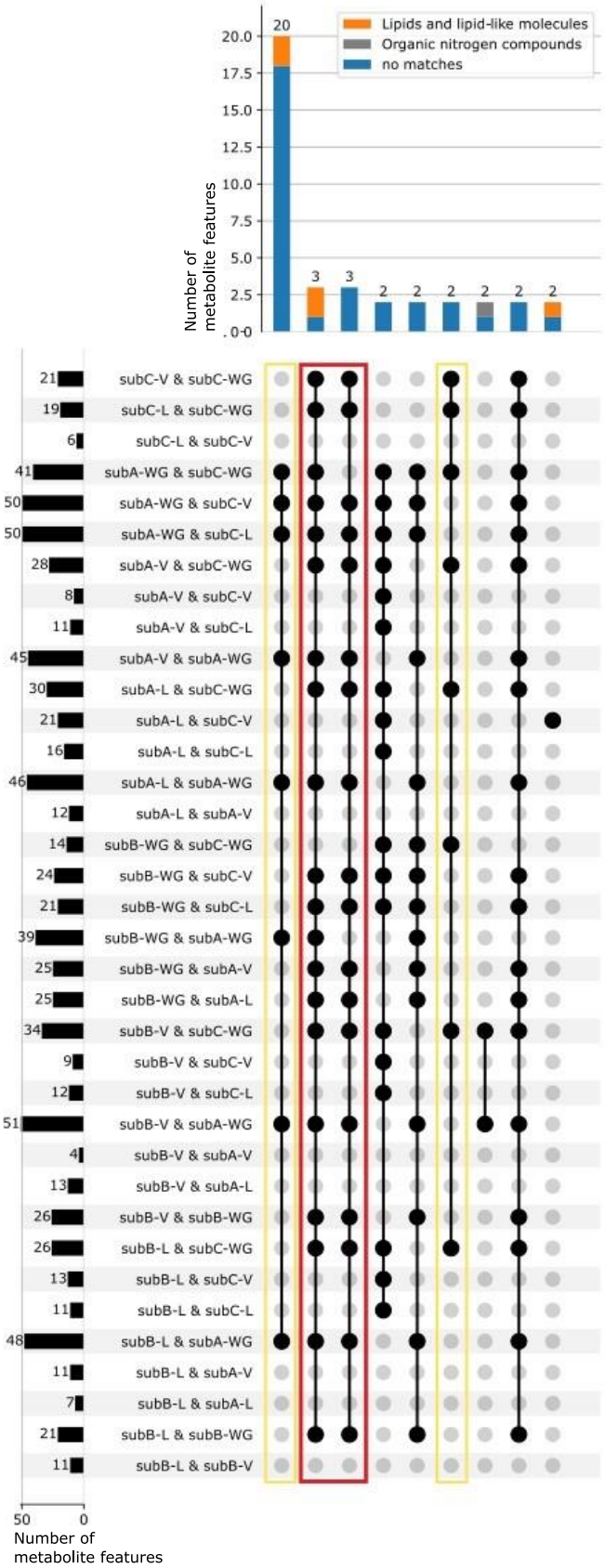

Lipids and lipid-like molecules
Organic nitrogen compounds
no matches
Number of metabolite features
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
20
3
3
2
2
2
2
2
2
21 subC-V & subC-WG
19 subC-L & subC-WG
6 subC-L & subC-V
41 subA-WG & subC-WG
50 subA-WG & subC-V
50 subA-WG & subC-L
28 subA-V & subC-WG
8 subA-V & subC-V
11 subA-V & subC-L
45 subA-V & subA-WG
30 subA-L & subC-WG
21 subA-L & subC-V
16 subA-L & subC-L
46 subA-L & subA-WG
12 subA-L & subA-V
14 subB-WG & subC-WG
24 subB-WG & subC-V
21 subB-WG & subC-L
39 subB-WG & subA-WG
25 subB-WG & subA-V
25 subB-WG & subA-L
34 subB-V & subC-WG
9 subB-V & subC-V
12 subB-V & subC-L
51 subB-V & subA-WG
4 subB-V & subA-V
13 subB-V & subA-L
26 subB-V & subB-WG
26 subB-L & subC-WG
13 subB-L & subC-V
11 subB-L & subC-L
48 subB-L & subA-WG
11 subB-L & subA-V
7 subB-L & subA-L
21 subB-L & subB-WG
11 subB-L & subB-V
50
0
Number of metabolite features

Figure S3.4: Unique and overlapping metabolite features at differential levels between donors and diets in serum (UpSet plot).
Kruskal-Wallis, FDR-corrected, $p<0.05$ and fold change ≥ 2. Figure subset to only include overlaps containing two or more features. Connected black circles represent experimental diet-donor pairs that had significantly modulated metabolite features in common (intersections between groups). Colored columns represent the number of microbiome-modulated metabolite features in each intersection, colored by Classyfire superclass. Row bars represent the number of metabolite features in a given set, for example 21 features that differed between SubC under a vegetable diet (SubC-V) and SubC under a whole grain diet (SubC-WG). Features whose differential abundance was characteristic of a given donor+diet combination is marked by gold boxes. Features whose differential abundance was characteristic of the whole grain diet are marked by a red box. V, vegetable; L, legume; WG, whole grain diets. Last row (subB-L & subB-V) is empty because only overlaps of >2 metabolite features are shown here.

Supplementary Material C

Author List and Affiliations

Jacob J. Haffner[1,2], Keith M. Prufer[3], Hannah V. Mattson[3], Kari L. Schleher[3,4,5], Roxanne Swentzell[6], Horvey Palacios[1,2], Nadia Neff[3], Jose Mes[7], Christina Warinner[8], Laura-Isobel McCall[2,9,10], Cecil M. Lewis, Jr.[1,2]

[1]Department of Anthropology, University of Oklahoma, Norman, OK, USA.

[2]Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, OK, USA.

[3]Department of Anthropology, University of New Mexico, Albuquerque, NM, USA.

[4]Maxwell Museum of Anthropology, University of New Mexico, Albuquerque, NM, USA.

[5]Crow Canyon Archaeological Center, Cortez, CO, USA.

[6]Flowering Tree Permaculture Institute, Española, NM, USA.

[7]Uchben'kaj Kin Ajaw Association, Santa Cruz, Belize.

[8]Department of Anthropology, Harvard University, Cambridge, MA, USA.

[9]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA.

[10]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA.

Author Contributions

C.M.L, T.W., and K.M.P. conceived and designed the study. C.W. led Mexico sample collection and handled sample treatment, including subsampling, and storage. J.J.H, H.V.M., and K.L.S. collected New Mexico samples under direction and guidance from R.S. J.J.H. led New Mexico sample treatment, including aliquoting and storage. K.M.P. and N.N. collected Southern Belize samples with direction and support from J.M. N.N. and H.P. led Southern Belize sample prep, including subsampling and storage. C.M.L. provided funding support to work with samples and L.-I.M. provided laboratory and LC-MS/MS resources. J.J.H. performed laboratory experimentation involving metabolite extraction and conducted LC-MS/MS under direction from L.-I.M. J.J.H. handled data analysis under guidance from L.-I.M. and C.M.L. Manuscript was written by J.J.H. with assistance from C.M.L.

Acknowledgements

We thank the Flowering Tree Permaculture Institute, Albuquerque Plants of the Southwest plant nursery, Oaxaca City Mercado Benito Juárez, Oaxaca Tlacolula market, Villahermosa market, Bladen Nature Reserve, and Santa Cruz Village Toledo district for their support and contributing samples for this project. We also thank the McCall Lab at OU for providing resources for laboratory and mass spectrometry analyses.

Supplementary Table 4.1

Table S4.1: Sampled plant/diet list.
"SO" represents Summer Oaxaca, "WO" represents Winter Oaxaca, "_F" represents a fruit sample, "_S" represents a seed sample, "SV" represents Summer Villahermosa, "NM.FT" represents New Mexico Flowering Tree, "NM.PSW" represents New Mexico Plants of the Southwest, and "SB" represents Southern Belize. "NA" values represent values not available.

SampleID	Common name	Taxonomy	Region	Site	Botanical Part	Year
SO01	Chile pulla	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO02	Chile de onza	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO03	Chile morita	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO04	Chile costeno amarillo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO05	Chile costeno	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO06	Chile chiltepec	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO07	Chile de arbol	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO08	Chipotle meco	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO09	Chile cascabe	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO10	Chile mulato	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO11	Chipote rojo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO12	Chile guajillo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006

SO13	Chile ancho rojo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO14	Pasilla mexicana	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO15	Chile tuxtla	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO16	Chile chintepe	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO17	Chile mulato	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO18	Chile ancho negro	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO19	Pasilla oaxaqueno	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO20	Chile de onza	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO21	Chile de agua	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO22	Chiuacle	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO23	Chile pico de gallo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO24	Chile chilhuacle rojo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO25	Chile de onza amarillo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO26	Chile costeno amarillo	*Capsicum annuum*	Mexico	Summer Oaxaca	Fruit	2006
SO28	Ciruela	*Spondias sp.*	Mexico	Summer Oaxaca	Fruit	2006
SO29	Maracuya	*Passiflora edulis*	Mexico	Summer Oaxaca	Fruit	2006
SO30	Guava	*Psidium guajava*	Mexico	Summer Oaxaca	Fruit	2006

SO31	Aguacate criollo	*Persea americana*	Mexico	Summer Oaxaca	Fruit	2006
SO32	Chayote grande	*Sechium edulis*	Mexico	Summer Oaxaca	Fruit	2006
SO33	Chayote	*Sechium edulis*	Mexico	Summer Oaxaca	Fruit	2006
SO34	Aguacate criollo	*Persea americana*	Mexico	Summer Oaxaca	Fruit	2006
SO35	Mesquite	*Prosopis sp.*	Mexico	Summer Oaxaca	Seed	2006
SO36	Nopal cacuts	*Opuntia ficus-indica*	Mexico	Summer Oaxaca	Stem	2006
SO37	Avocado	*Persea americana*	Mexico	Summer Oaxaca	Fruit	2006
SO38	Red prickly pear	*Opuntia ficus-indica*	Mexico	Summer Oaxaca	Fruit	2006
SO39	Green prickly pear	*Opuntia ficus-indica*	Mexico	Summer Oaxaca	Fruit	2006
SO41	Maguey worm	*Hypopta agavis*	Mexico	Summer Oaxaca	Insect	2006
SO42	Chapulines pequeno	*Sphenarium spp.*	Mexico	Summer Oaxaca	Insect	2006
SO43	Chapulines grande	*Sphenarium spp.*	Mexico	Summer Oaxaca	Insect	2006
WO02	Cola de caballo	*Equisetum spp.*	Mexico	Winter Oaxaca	Stem	2005
WO03	Gordolobo	*Gnaphalium spp.*	Mexico	Winter Oaxaca	Flower	2005
WO04	Cascara de encino rojo	*Quercus rubra*	Mexico	Winter Oaxaca	Wood	2005
WO08	Cascara de pallo dulce	*Eysenhardtia polystachya*	Mexico	Winter Oaxaca	Wood	2005
WO10	Pinguica	*Arctostaphylos pungens*	Mexico	Winter Oaxaca	Leaf	2005
WO12	Guaje	*Leucaena esculenta*	Mexico	Winter Oaxaca	Seed	2005

WO13	Nopal cactus	*Opuntia ficus-indica*	Mexico	Winter Oaxaca	Stem	2005
WO14	Sweet potato	*Ipomaea batatas*	Mexico	Winter Oaxaca	Tuber	2005
WO15	Chayote (smooth)	*Sechium edulis*	Mexico	Winter Oaxaca	Fruit	2005
WO16	Tomatillo	*Physalis philadelphica*	Mexico	Winter Oaxaca	Fruit	2005
WO17	Habanjero	*Capsicum chinense*	Mexico	Winter Oaxaca	Fruit	2005
WO18	Jalapeno	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO19	Serrano pepper	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO20	Jalapeno	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO22	Huiche squash	*Cucurbita pepo*	Mexico	Winter Oaxaca	Seed	2005
WO23	Tambulo squash	*Cucurbita pepo*	Mexico	Winter Oaxaca	Seed	2005
WO24	Huiche squash	*Cucurbita pepo*	Mexico	Winter Oaxaca	Fruit	2005
WO25	Huiche squash	*Cucurbita pepo*	Mexico	Winter Oaxaca	Flower	2005
WO26	Chipil	*Crotelaria sp.*	Mexico	Winter Oaxaca	Leaf	2005
WO27	Maize	*Zea mays*	Mexico	Winter Oaxaca	Seed	2005
WO28	Berro	*Nasturtium officinale*	Mexico	Winter Oaxaca	Leaf	2005
WO29	Guatomate	*Lycopersicon esculentum*	Mexico	Winter Oaxaca	Fruit	2005
WO31	Zucchini	*Cucurbita pepo*	Mexico	Winter Oaxaca	Fruit	2005
WO34	Jitomate rojo	*Lycopersicon esculentum*	Mexico	Winter Oaxaca	Fruit	2005
WO35	Jitomate	*Lycopersicon esculentum*	Mexico	Winter Oaxaca	Fruit	2005
WO36	Guava	*Psidium guajava*	Mexico	Winter Oaxaca	Fruit	2005

WO37	Cacao	*Theobroma cacao*	Mexico	Winter Oaxaca	Seed	2005
WO40	Jicama	*Pachyrhizus erosus*	Mexico	Winter Oaxaca	Tuber	2005
WO41	Aguacate criollo	*Persea americana*	Mexico	Winter Oaxaca	Fruit	2005
WO42	Nance	*Byrsonima crassifolia*	Mexico	Winter Oaxaca	Fruit	2005
WO43	Guanabana	*Annona muricata*	Mexico	Winter Oaxaca	Fruit	2005
WO44	Mamey	*Pouteria sapota*	Mexico	Winter Oaxaca	Fruit	2005
WO45	Ciruela	*Spondias sp.*	Mexico	Winter Oaxaca	Fruit	2005
WO46	Chicozapote	*Manilkara zapotilla*	Mexico	Winter Oaxaca	Fruit	2005
WO47	Zapote negro	*Diospyros digyna*	Mexico	Winter Oaxaca	Fruit	2005
WO48	Zapote negro (ripe)	*Diospyros digyna*	Mexico	Winter Oaxaca	Fruit	2005
WO49	Black walnut	*Juglans sp.*	Mexico	Winter Oaxaca	Seed	2005
WO50	Bean (flor de mayo)	*Phaseolus vulgaris*	Mexico	Winter Oaxaca	Seed	2005
WO51	Bean (uallo)	*Phaseolus vulgaris*	Mexico	Winter Oaxaca	Seed	2005
WO52	Bean (blanco)	*Phaseolus vulgaris*	Mexico	Winter Oaxaca	Seed	2005
WO73_F	Yellow squash	*Cucurbita sp.*	Mexico	Winter Oaxaca	Fruit	2005
WO73_S	Yellow squash	*Cucurbita sp.*	Mexico	Winter Oaxaca	Seed	2005
WO74_F	White squash	*Cucurbita moschata*	Mexico	Winter Oaxaca	Fruit	2005
WO74_S	White squash	*Cucurbita moschata*	Mexico	Winter Oaxaca	Seed	2005
WO77	Chapulines grande	*Sphenarium spp.*	Mexico	Winter Oaxaca	Meat	2005
WO83	Pulque	*Agave spp.*	Mexico	Winter Oaxaca	Alcohol	2005
WO84	Amaranth	*Amaranthus spp.*	Mexico	Winter Oaxaca	Seed cake	2005
WO86	Chipil	*Crotelaria sp.*	Mexico	Winter Oaxaca	Leaf	2005

WO88	Camote	*Ipomaea batatas*	Mexico	Winter Oaxaca	Tuber	2005
WO89	Pimento	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO90	Yuca	*Manihot esculenta*	Mexico	Winter Oaxaca	Tuber	2005
WO91	Mexican oregano	*Lippia graveolens*	Mexico	Winter Oaxaca	Leaf	2005
WO93	Tepeguaje	*Leucaena sp.*	Mexico	Winter Oaxaca	Leaf	2005
WO94	Achiote	*Bixa orellana*	Mexico	Winter Oaxaca	Seed	2005
WO98	Mesquite	*Prosopis sp.*	Mexico	Winter Oaxaca	Pod/seed	2005
WO100	Chayote (spiny)	*Sechium edulis*	Mexico	Winter Oaxaca	Fruit	2005
WO101	Chicozapote	*Manilkara zapotilla*	Mexico	Winter Oaxaca	Fruit	2005
WO102	Chile guajillo	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO103	Chile	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO104	Chile ancho	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO105	Chile ancho	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO106	Chile	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO108	Chile de arbol	*Capsicum annuum*	Mexico	Winter Oaxaca	Fruit	2005
WO110	Epazote	*Chenopodium ambrosioides*	Mexico	Winter Oaxaca	Leaf	2005
SV01	Nance	*Byrsonima crassifolia*	Mexico	Summer Villahermosa	Fruit	2006
SV02	Papaya	*Carica papaya*	Mexico	Summer Villahermosa	Fruit	2006
SV03	Guava	*Psidium guajava*	Mexico	Summer Villahermosa	Fruit	2006
SV06	Avocado	*Persea americana*	Mexico	Summer Villahermosa	Fruit	2006
SV07	Chile picante	*Capsicum annuum*	Mexico	Summer Villahermosa	Fruit	2006
SV08	Chile dulce	*Capsicum annuum*	Mexico	Summer Villahermosa	Fruit	2006
SV09	Chile habañero	*Capsicum chinense*	Mexico	Summer Villahermosa	Fruit	2006

SV10	Yuca	*Manihot esculenta*	Mexico	Summer Villahermosa	Tuber	2006
SV11	Chayote	*Sechium edulis*	Mexico	Summer Villahermosa	Fruit	2006
SV13	Squash	*Cucurbita sp.*	Mexico	Summer Villahermosa	Fruit	2006
SV14	Pitaya	*Hylocereus undatus*	Mexico	Summer Villahermosa	Fruit	2006
SV15	Guajilote	*Parmentiera edulis*	Mexico	Summer Villahermosa	Fruit	2006
SV18	Nopal cactus	*Opuntia ficus-indica*	Mexico	Summer Villahermosa	Stem	2006
SV19	Jicama	*Pachyrhizus erosus*	Mexico	Summer Villahermosa	Tuber	2006
SV23	Jitomate rojo	*Lycopersicon esculentum*	Mexico	Summer Villahermosa	Fruit	2006
SV24	Jitomate verde	*Lycopersicon esculentum*	Mexico	Summer Villahermosa	Fruit	2006
SV25	Green prickly pear	*Opuntia ficus-indica*	Mexico	Summer Villahermosa	Fruit	2006
SV26	Chile	*Capsicum annuum*	Mexico	Summer Villahermosa	Fruit	2006
SV28	Shrimp	*NA*	Mexico	Summer Villahermosa	Shrimp	2006
SV30	Chiquijul	*Bactris sp.*	Mexico	Summer Villahermosa	Fruit	2006
SV31	Maize	*Zea mays*	Mexico	Summer Villahermosa	Seed	2006
SV32	Pineapple	*Ananas cosmosus*	Mexico	Summer Villahermosa	Fruit	2006
SV37	Cancerillo	*Asclepias curassavica*	Mexico	Summer Villahermosa	Leaf	2006
SV38	Maize (blanco)	*Zea mays*	Mexico	Summer Villahermosa	Seed	2006
SV39	Bean (bayo)	*Phaseolus vulgaris*	Mexico	Summer Villahermosa	Seed	2006
SV40	Bean (flor de mayo)	*Phaseolus vulgaris*	Mexico	Summer Villahermosa	Seed	2006
SV42	Pipian squash	*Cucurbita mixta*	Mexico	Summer Villahermosa	Seed	2006
SV45	Sunflower	*Helianthus annuus*	Mexico	Summer Villahermosa	Seed	2006
SV55	Pinguica	*Arctosaphylos pungens*	Mexico	Summer Villahermosa	Fruit	2006

SV56	Toronjil	*Agastache mexicana*	Mexico	Summer Villahermosa	Leaf	2006
SV57	Bean (negro)	*Phaseolus vulgaris*	Mexico	Summer Villahermosa	Seed	2006
SV58	Bean (blanco)	*Phaseolus vulgaris*	Mexico	Summer Villahermosa	Seed	2006
SV59	Squash	*Cucurbita sp.*	Mexico	Summer Villahermosa	Seed	2006
SV60	Turkey (cooked)	*Meleagris gallopavo*	Mexico	Summer Villahermosa	Bird	2006
SV61	Chaya	*Cnidoscolus chayamansa*	Mexico	Summer Villahermosa	Leaf	2006
SV62	Chile anachito	*Capsicum annuum*	Mexico	Summer Villahermosa	Fruit	2006
SV65	Chipil	*Crotelaria sp.*	Mexico	Summer Villahermosa	Leaf	2006
SV67	Epazote	*Chenopodium ambrosioides*	Mexico	Summer Villahermosa	Leaf	2006
SV68	Melocoton, cassabanana	*Sicana odorifera*	Mexico	Summer Villahermosa	Fruit	2006
SV70	Bandesopa, poshte	*Annona scleroderma*	Mexico	Summer Villahermosa	Fruit	2006
SV71	Macal/Cocoyam	*Xanthosoma sagittifolium*	Mexico	Summer Villahermosa	Tuber	2006
SV73	Sweet potato	*Ipomaea batatas*	Mexico	Summer Villahermosa	Tuber	2006
SV74	Maracuya	*Passiflora edulis*	Mexico	Summer Villahermosa	Fruit	2006
SV75	Guava	*Psidium guajava*	Mexico	Summer Villahermosa	Fruit	2006
SV76	Mamey	*Pouteria sapota*	Mexico	Summer Villahermosa	Fruit	2006
SV77	Chicozapote	*Manilkara zapotilla*	Mexico	Summer Villahermosa	Fruit	2006
SV78	Honey (bee)	NA	Mexico	Summer Villahermosa	Honey	2006
SV79	Crab	NA	Mexico	Summer Villahermosa	Meat	2006
SV80	Shrimp	NA	Mexico	Summer Villahermosa	Meat	2006
SV81	Crayfish	NA	Mexico	Summer Villahermosa	Meat	2006

NM_FT01	Amaranth (red)	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT03	Amaranth (red)	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT06	Amaranth (red)	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT08	Wild spinach	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT15	Wild spinach	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT16	Mesquite	NA	New Mexico	Flowering Tree	Ground-up	2022
NM_FT18	Indian rice	NA	New Mexico	Flowering Tree	Grain	2022
NM_FT20	Rocky mountain beeweed	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT23	Rocky mountain beeweed	NA	New Mexico	Flowering Tree	Flower	2022
NM_FT24	Purslane	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT26	Purslane	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT28	Dandelion	NA	New Mexico	Flowering Tree	Flower	2022
NM_FT30	Dandelion	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT31	Dandelion	NA	New Mexico	Flowering Tree	Stem	2022

NM_FT32	Pigweed	NA	New Mexico	Flowering Tree	Spiny leaf	2022
NM_FT34	Pigweed	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT36	Pigweed	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT37	Artichoke	NA	New Mexico	Flowering Tree	Root	2022
NM_FT39	Tomatillo (round)	NA	New Mexico	Flowering Tree	Bulb	2022
NM_FT40	Tomatillo (round)	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT42	Tomatillo (long)	NA	New Mexico	Flowering Tree	Bulb	2022
NM_FT43	Tomatillo (long)	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT44	Tomatillo (long)	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT45	Cushaw	NA	New Mexico	Flowering Tree	Flesh	2022
NM_FT46	Cushaw	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT47	Dried squash	NA	New Mexico	Flowering Tree	Flesh	2022
NM_FT49	Tewa brown-seeded watermelon	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT50	Ancient watermelon	NA	New Mexico	Flowering Tree	Seed	2022

NM_FT51	Santo Domingo watermelon	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT52	Taos Pueblo Squash	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT53	Pueblo Squash	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT54	Sand cherry	NA	New Mexico	Flowering Tree	Fruit	2022
NM_FT55	Sand cherry	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT56	Sand cherry	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT57	Cultivated raspberry	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT59	Cultivated currants	NA	New Mexico	Flowering Tree	Berry	2022
NM_FT61	Wild plum	NA	New Mexico	Flowering Tree	Fruit (purple)	2022
NM_FT62	Wild plum	NA	New Mexico	Flowering Tree	Fruit (yellow)	2022
NM_FT63	Choke cherry (jam possibly with honey)	NA	New Mexico	Flowering Tree	Jam	2022
NM_FT64	Prickly pear	NA	New Mexico	Flowering Tree	Fruit	2022
NM_FT66	Cholla	NA	New Mexico	Flowering Tree	Bud	2022
NM_FT67	Pinyon	NA	New Mexico	Flowering Tree	Berry	2022

NM_FT69	Wild currant	NA	New Mexico	Flowering Tree	Berry	2022
NM_FT71	Pueblo tobacco	NA	New Mexico	Flowering Tree	Leaf (dried)	2022
NM_FT73	Pueblo tobacco	NA	New Mexico	Flowering Tree	Stem (dried)	2022
NM_FT74	Sumac	NA	New Mexico	Flowering Tree	Bulb	2022
NM_FT75	Large sunflower	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT76	Small sunflower	NA	New Mexico	Flowering Tree	Flower	2022
NM_FT77	Small sunflower	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT78	Small sunflower	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT83	Cota/indian tea	NA	New Mexico	Flowering Tree	Flower (dried)	2022
NM_FT85	Cota/indian tea	NA	New Mexico	Flowering Tree	Stem (dried)	2022
NM_FT86	Devil's claw (green)	NA	New Mexico	Flowering Tree	Bulb	2022
NM_FT87	Devil's claw (dried black)	NA	New Mexico	Flowering Tree	Bulb	2022
NM_FT88	Juniper	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT90	Juniper	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT91	Mormon tea	NA	New Mexico	Flowering Tree	Leaf	2022

NM_FT93	Mormon tea	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT95	Salt bush	NA	New Mexico	Flowering Tree	Leaf	2022
NM_FT96	Salt bush	NA	New Mexico	Flowering Tree	Stem	2022
NM_FT97	Squash (green)	NA	New Mexico	Flowering Tree	Flesh	2022
NM_PSW99	Scarlet globemallow	NA	New Mexico	Plants of the Southwest	Petals	2022
NM_PSW100	Texas ranger heavenly cloud	NA	New Mexico	Plants of the Southwest	Petals	2022
NM_PSW101	Small leaved sumac	NA	New Mexico	Plants of the Southwest	Fruit	2022
NM_PSW102	Sweet sand verbena	NA	New Mexico	Plants of the Southwest	Petals	2022
NM_PSW103	New Mexico privet	NA	New Mexico	Plants of the Southwest	Leaf	2022
NM_FT105	Chimayo	NA	New Mexico	Flowering Tree	Flesh	2022
NM_FT106	Chimayo	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT108	Corn (blue)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT109	Corn (bright yellow)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT110	Corn (dried dark yellow)	NA	New Mexico	Flowering Tree	Kernel	2022

NM_FT111	Corn (dried white-pink)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT112	Corn (dried red)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT113	Corn (dried black-purple)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT114	Corn (dried white)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT115	Yellow pasole (processed with lime)	NA	New Mexico	Flowering Tree	Kernel	2022
NM_FT116	Blue atole (roasted with lime)	NA	New Mexico	Flowering Tree	Powder	2022
NM_FT117	Red masa (processed with commercial lime)	NA	New Mexico	Flowering Tree	Powder	2022
NM_FT118	Red posole (processed with lime)	NA	New Mexico	Flowering Tree	Powder	2022
NM_FT119	Flowering tree chile	NA	New Mexico	Flowering Tree	Flesh	2022
NM_FT120	Flowering tree chile	NA	New Mexico	Flowering Tree	Seed	2022
NM_FT121	Yellow bean	NA	New Mexico	Flowering Tree	Bean	2022

NM_FT122	Taos read bean	NA	New Mexico	Flowering Tree	Bean	2022
NM_FT123	Pueblo black bean	NA	New Mexico	Flowering Tree	Bean	2022
NM_FT124	Anasazi bean	NA	New Mexico	Flowering Tree	Bean	2022
NM_FT125	White tepary bean	NA	New Mexico	Flowering Tree	Bean	2022
NM_FT126	Pueblo pinto bean	NA	New Mexico	Flowering Tree	Bean	2022
NM_FT127	Soil	NA	New Mexico	Flowering Tree	Soil	2022
SB01	Toilet paper leaf	NA	Southern Belize	Toledo district	NA	2021
SB02	Jippy Jappa	NA	Southern Belize	Toledo district	NA	2021
SB03	Chaya	NA	Southern Belize	NA	NA	2022
SB04	Small pepper	NA	Southern Belize	Toledo district	NA	2022
SB05	Pumpkin	NA	Southern Belize	Toledo district	NA	2021
SB06	White cacao	NA	Southern Belize	Bladen Nature Reserve	NA	2022
SB07	Cassava	NA	Southern Belize	Toledo district	NA	2021
SB08	Coco yam	NA	Southern Belize	Toledo district	NA	2021
SB09	Inga	NA	Southern Belize	Toledo district	NA	2021

SB11	Saw Palmetto	NA	Southern Belize	Toledo district	NA	2021
SB12	Pacaya	NA	Southern Belize	Toledo district	NA	2021
SB13	Wild passion fruit	NA	Southern Belize	Bladen Nature Reserve	NA	2022
SB14	Kulah, Jippy Jappa	NA	Southern Belize	Toledo district	NA	2021
SB15	Chiki Ginger, Heleconia	NA	Southern Belize	Toledo district	NA	2021
SB16	Cat's Balls	NA	Southern Belize	Toledo district	NA	2021
SB17	Heleconia	NA	Southern Belize	Toledo district	NA	2021
SB18	Cassava	NA	Southern Belize	Toledo district	NA	2021
SB19	Cohune nut	NA	Southern Belize	Toledo district	NA	2021
SB20	Cacao	NA	Southern Belize	Toledo district	NA	2021
SB21	White cacao	NA	Southern Belize	Toledo district	NA	2021
SB22	Pumpkin	NA	Southern Belize	Toledo district	NA	2021
SB23	Cohune nut	NA	Southern Belize	Toledo district	Nut	2022
SB24	Red bean	NA	Southern Belize	Toledo district	Bean	2021

SB25	Black bean	NA	Southern Belize	Toledo district	Bean	2021
SB26	Red corn	NA	Southern Belize	Toledo district	Kernel	2021
SB27	Purple corn	NA	Southern Belize	Toledo district	Kernel	2021
SB28	White corn	NA	Southern Belize	Toledo district	Kernel	2021
SB29	Yellow corn	NA	Southern Belize	Toledo district	Kernel	2021
SB30	Yellow corn	NA	Southern Belize	Toledo district	Kernel	2021
SB31	White corn	NA	Southern Belize	Toledo district	Kernel	2021
SB32	Purple corn	NA	Southern Belize	Toledo district	Kernel	2021
SB33	Yellow corn	NA	Southern Belize	Toledo district	Kernel	2022
SB34	White corn	NA	Southern Belize	Toledo district	Kernel	2022
SB35	Purple corn	NA	Southern Belize	Toledo district	Kernel	2022
ApisA	European honeybee	NA	NA	NA	Honey	2022
ApisB	European honeybee	NA	NA	NA	Honey	2022
MelA	Melipona bee	NA	NA	NA	Honey	2022
MelB	Melipona bee	NA	NA	NA	Honey	2022